Universitat Politècnica de Catalunya
Facultat de Matemàtiques i Estadística

Degree in Mathematics

Bachelor's Degree Thesis

Effects of a professional development program on GTA teaching effectiveness

Joan Espar Masip

Supervised by Emily Alicea-Muñoz, Mike Schatz, Pedro Delicado

January, 2017

Thanks to the Universitat Politècnica de Catalunya, specially to the CFIS to give me the opportunity to go to Georgia Tech to do my final degree work. In concrete many thanks to Miguel Ángel Barja for helping me with all the process, and many thanks also to Pedro Delicado for being my advisor at the UPC.

On the other hand, I am very grateful to all the Physics Education Research Group for allowing me to work with them, helping me with all the problems and all the work I have done and also for teaching me how is the real life of a research group. Many thanks to Mike Schatz for accepting and allowing me to go to Georgia Tech and also for all the advice which has helped me a lot.

Finally, I am so much more than grateful to Emily Alicea-Muñoz who not only has been my advisor, but also has taught me almost everything I have needed to do my work and has cared for me during the four months I have been there.

Abstract

For the last three years, the School of Physics at Georgia Tech has been preparing new Graduate Teaching Assistants (GTAs) with a mentoring and development program that focuses on pedagogy, physics content, and professional development strategies. Our goal is to produce effective GTAs who have a positive impact on student learning, while honing the skills they need to succeed in their future careers. We want to determine the program's impact on GTAs' overall teaching effectiveness as well as their performance in some important aspects of a proper teaching. To do that we performed several statistical analyses of students' responses to end-of-semester GTA evaluations. Here we present the results of our analyses, in particular the comparison between GTAs who participated in the program and GTAs before the program went into effect, considering also other variables that could affect to the evaluations results.

Keywords

GTA, teaching, assistant, effectiveness, statistics, physics, education

Contents

Part I
Introduction

1. Background and Motivation

Graduate students working as Teaching Assistants (GTAs) perform a variety of duties, such as teaching lab or recitation classes, grading, or tutoring students. They are especially important in the teaching of large-enrollment introductory physics courses, where students spend approximately half of their in-class time supervised by GTAs. Moreover, the influence of the GTAs in the first years of the students' education can be crucial for their academic and professional future.

One of the earliest proposals for physics TA training was in a conference held at Lake Wilderness, Washington, in 1969 [15]. In this conference, and many posterior ones, there was noticed an increased concern for the improvement of the graduate preparation for people who wish to teach physics in college and hence, there were proposed and discussed several training programs for students to prepare them to teach.

One of these preparation courses took place in the Ohio University and it had three main goals: to increase the interest of the graduate students in physics education, give them useful information for when they are teaching, and the opportunity to teach with peer evaluation. At the end of the training course, the graduate students evaluated it, and the results were significantly good, as the graduate students found it instructive and useful, proving that this type of course can be beneficial for the graduate students and, therefore, for the students whom the graduate students will be teaching [15].

Another course was implemented in the University of Missouri-Columbia in 1973 [12]. The principal objectives of this course was to improve the quality of the undergraduate physics laboratories and improve the teaching ability of the graduate teaching assistants. After the semester the graduate students had acquired new teaching knowledge, which were able to apply to their classes, and the overall result was positive, both for graduate and undergraduate students.

Many other preparation courses were developed in different universities during the following years. The majority of them were focused on the same aspect as the two mentioned here: providing the graduate students better knowledge about teaching; which includes more physic content, grading techniques, classroom management and in definitive, everything that could improve their teaching skills.

In the School of Physics at Georgia Tech, first-time GTAs usually help professors with the Introductory Physics I and II courses (introductory mechanics and electromagnetism respectively). These two courses are usually taken by students in their first or second year of university, so the effect of the teacher, professor or GTA, becomes more relevant. For that reason, it's important to give the GTAs appropriate preparation so they can teach, grade and help students more efficiently, especially during their first semester as a GTA.

Before 2010 the preparation for the new GTAs consisted only of a one- or two-day orientation where they were taught about institute policies and grading, as well as their duties and responsibilities. They also had weekly meetings with their coordinator to discuss the following week assignments. Between 2010 and 2013, new GTAs received additional preparation in the form of seminars offered by the Center for Teaching and Learning. In 2013, the department began a formal GTA preparation program structured as a one-semester course, CETL 8000 PH. This training program is a more exhaustive preparation for the GTAs based on pedagogy, physics content and professional development strategies, and one of its main

objectives is to give the GTAs not only the tools they need to become better teachers, but also to improve their future research careers.

Three years after the implementation of CETL 8000 PH, our goal is to discuss its effect on the GTAs' teaching skills. To do that, we will analyze the results of the end-of-semester student evaluation surveys. We will focus on the results for the GTAs' first year of teaching, as this is when the effect of the training program should be more noticeable, and then also see the evolution of the GTAs over time. Finally, we want to determine in which aspects of the GTAs the program has had a bigger impact, and use this information to improve the program in the following years.

The importance of these preparation programs was crucial as, like [6] say, graduate teaching assistants are apprehensive about teaching for the first time. First-time GTAs have a lot of responsibilities apart from their teaching obligations, such as their own classes and their research, and adding to those a self-learning how to teach would probably lead to a bad teaching experience and a worst learning from their undergraduate students. Moreover, a proper training course would not only provide the abilities to become a better teacher, but to become a better professional in their own fields [4]. Although it was well known that the preparation for first-time graduate teaching assistants was beneficial both for them and for the undergraduate students they teach, there does not exist a universal TA training program. Instead, TA training varies from institution to institution, and in some cases it continues to be nonexistent.

2. GTAs at Georgia Tech School of Physics

2.1 GTA duties and responsibilities

First-time GTAs: Intro Physics (M&I / Traditional)

As we have said before, Graduate Teaching Assistants (GTAs) have to perform a variety of duties such as teaching lab or recitation classes, grading, or tutoring students. Most graduate students work as TAs in their first year of grad school, and the majority of these are assigned to teach for the introductory physics classes. We will be analyzing student evaluations of GTA performance only for these courses: PHYS 2211, Intro Physics I (mechanics) and PHYS 2212, Intro Physics II (electromagnetism). It is important to note that at Georgia Tech there are two 'flavors' of introductory physics – Traditional and Matter & Interactions (M&I) – and the GTA duties vary somewhat between the two flavors.

The Traditional course uses the textbook Physics for Scientists and Engineers: A Strategic Approach (3rd Edition) by Randall D. Knight [7], and the GTAs spend 2 hours of laboratory and 1 hour of recitation per week with the students (with different GTAs for labs and recitations). The course is developed with a traditional approach: for the Intro I they start with kinematics and Newton's laws, and the Intro II starts with the Gauss law and then both continue following a chronological structure. This approach is the one that has been usually taught in many different schools. In the laboratory classes they do classical experiments with equipment, such as calculating the gravity constant. The experiments are guided and the GTAs only have to help the students with their troubles while doing the experiments.

The M&I course has a different textbook: Matter and Interactions (4th Edition) by Ruth Chabay and Bruce Sherwood [2]. The GTAs only have laboratory classes 3 hours per week, and those lessons are very different than the Traditional ones. The majority fo the experiments are realized using Python and there is not as much equipment needed as in the Traditional labs. The development of the experiments it is also different as the GTAs start the problems and then the students have to continue by their own (the GTAs help them if it is needed).

Returning GTAs

After their first year in grad school, some grad students work as research assistants and some continue working as GTAs. These we refer to as Returning GTAs, since by that point they have worked as TAs for at least one year. Most Returning GTAs are assigned as graders for the upper-division and graduate classes, and only a few are assigned to the introductory classes. Figure 1 shows the distribution of GTA assignments for all graduate students enrolled in Spring 2016. We can clearly see there that most first-time GTAs are assigned to the intro classes.

2.2 GTA training before 2013

The graduate students of the School of Physics at Georgia Tech didn't receive any specific TA training before 2010, and it wasn't until 2013 that they started with a proper development program for new physics GTAs.

Before 2010, the new GTAs only had a one- or two-day orientation at the beginning of the semester. In that orientation they were taught about institute policies, their duties and responsibilities, and some

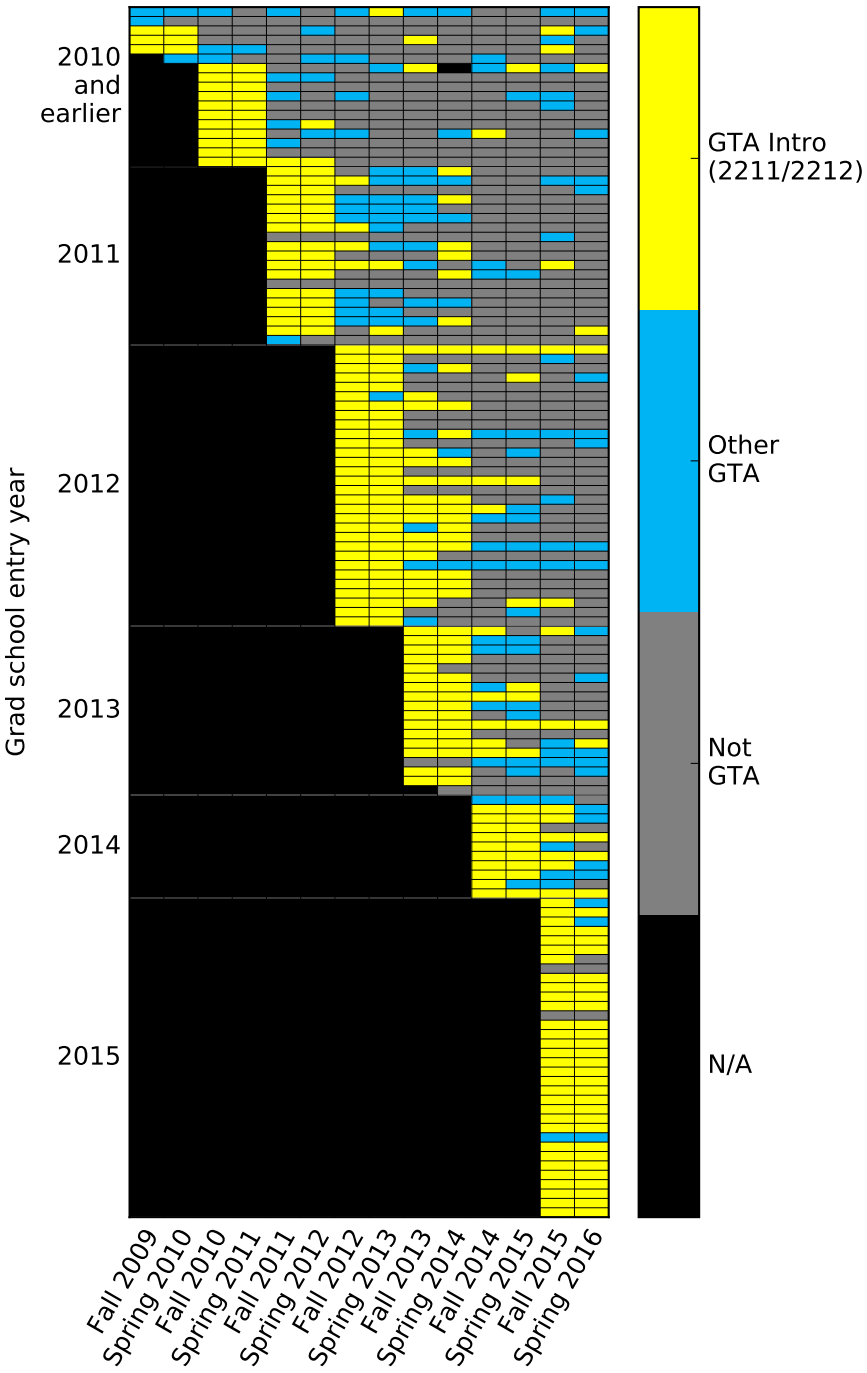


Figure 1: Visualization of the GTAs assigments by semesters. Each row is a graduate student, grouped by entry year. Figure by E. Alicea-Muñoz.

grading techniques. This orientation was complemented with weekly meetings with their coordinator to prepare the next week assignments. There wasn't any concrete information about teaching techniques or any improvement in physics content; we couldn't qualify this formation as a training program for new physics GTAs, as it only gave a basic introduction about their job. This lack of knowledge about teaching is one of the biggest concerns of the new GTAs and it directly affects their performance [18].

The first step in order to improve GTA training was done between 2010 and 2012. There wasn't a well defined training program but the new GTAs received more useful information about teaching. First of all, they had a general campus-wide orientation (New TA Orientation, or NTAO), where new GTAs from across the institute were taught about pedagogical subjects and institute policies. The new GTAs also had two meetings with the introductory physics coordinators, where they covered GTA duties and responsibilities, grading and other topics. Both the NTAO and the meetings were done the week before of the start of the semester. In addition to these, the new GTAs continued to have weekly meetings with their coordinator to discuss the topics of the following week classes. Finally, they had to attend to four pedagogical seminars offered by the Center for Teaching and Learning, which took place during the first two months of the semester. These seminars, along with the pedagogical content seen in the NTAO, were the start of getting some teaching knowledge which, according to previous experiences in other universities, was one of the most important things that a new GTA should know before they start teaching.

Although this new structure was better than the previous one, it wasn't a well-defined program which could cover all the important topics that a new GTA should learn. The pedagogical instruction received during this preparation was common to all the new GTAs of the university and, therefore, there wasn't a specific focus for those courses that the GTAs would have to teach. Moreover, the pedagogical seminars weren't very helpful and the new GTAs thought that the pedagogical information provided was not very useful or relevant for their teaching activities.

2.3 CETL 8000 PH

To improve the preparation of new GTAs, the School of Physics began a GTA development program in 2013. This program consists of a required one-credit course (CETL 8000 PH) for all first-year Ph.D. students, offered every year during the Fall semester. The course contents have changed during the three years that is has been running, but the main objectives and the core of the program have remained the same. The program was built on three main bases: Pedagogy, Physics Content and Professional Development. The principal difference between this new program and all the previous physics GTA training efforts at Georgia Tech is that this was the first comprehensive program developed specifically to prepare physics GTAs, and that should help them better with the concrete courses that they would teach.

Cycle 1

The first iteration of CETL 8000 PH was an adaptation of another training program developed by the Center for Teaching and Learning for the Georgia Tech School of Biology, but the contents were changed to make it more suitable for physics GTAs. It was done in Fall 2013 and it was divided into two parts: JumpStart to Teaching and the Semester Meetings.

The JumpStart happened before the start of the semester and covered the things GTAs should know before they start teaching. It included:

- Active learning: Discussions of active learning teaching methods and different learning styles.

- Engaging explanations: Introduction to the learner-centered teaching [5].
- Time management strategies.
- Microteaching: Peer evaluation of a short lesson by the other GTAs in the course.
- Classroom management: Discussion of students' motivations, and strategies to deal with problematic behaviors in the classroom.

During this first part, there also was a meeting with other experienced GTAs from the School of Physics. During this meeting the new GTAs could have some first-hand answers about their concerns and receive some advice from people that had been in their same situation.

The second part of the training was done during the semester. This consisted of one-hour meetings every two or three weeks, to discuss and improve some other aspects of their preparation. It included:

- Group work: Discussion of strategies to facilitate an effective group work.
- Grading: Getting practice in grading, being capable of give useful feedback to the students, strategies to deal with students complaining about their grades.
- Teaching what you don't know: Strategies to deal with situations where the GTAs are not able to answer a determinate question.
- Midterm evaluations: The GTAs would get feedback from their students in order to improve their teaching.
- Professional development: Introduction to the idea of teaching philosophy.

Cycle 2

This second edition of the training course was done in Fall 2014. The structure explained previously (JumpStart and Semester Meetings) remained more or less the same and a new element was added: Classroom Observations.

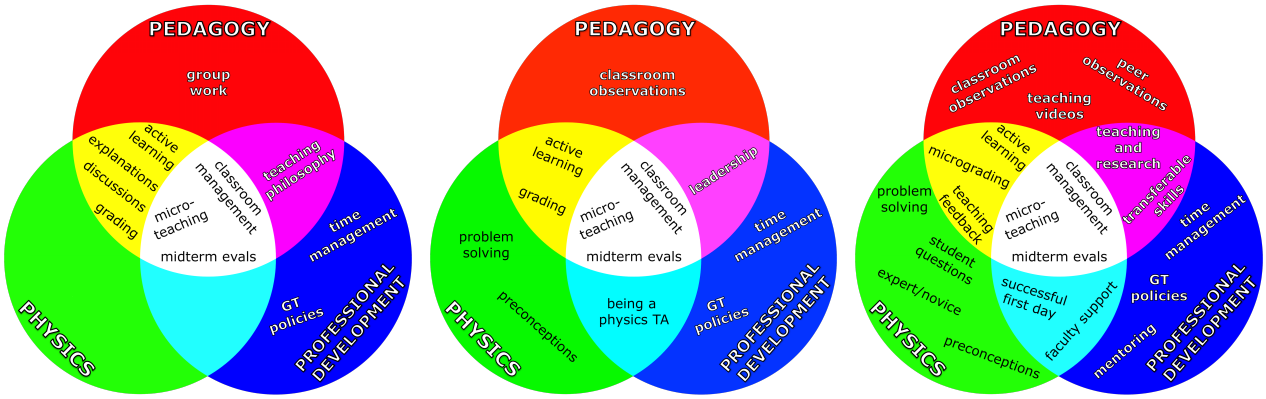
More physics content was added to the JumpStart lessons, and a new lesson on GTA video evaluation was also added to the Semester Meetings. In this lesson, the new GTAs watched video recording of experienced GTAs and then had to evaluate them and discuss the teaching strategies learned during the preparation.

The Classroom Observations was done by the CETL 8000 PH instructor/developer, who herself was an experienced GTA. She observed different classes given by the GTAs and determined if they were using the pedagogical techniques they had learned during the training course, and then she gave them feedback so they could improve their performance properly.

Cycle 3

The program continued in Fall 2015, with some new curriculum changes to make the course more robust and comprehensive. One such change was to increase the amount of feedback on GTAs' teaching and more discussions about their day-to-day feelings, giving the GTAs the opportunity to talk about their concerns and teaching questions more openly in a group setting. Another change was the inclusion of peer observations,

in addition to the regular classroom observations. In the regular classroom observations, the GTAs were observed by instructors (experienced GTAs) who then gave them feedback; in the peer observations, the new GTAs observed each other and gave each other feedback on their teaching.



(a) Cycle 1 (Fall 2013). (b) Cycle 2 (Fall 2014). (c) Cycle 3 (Fall 2015).

Figure 2: Evolution of the CETL 8000 PH curriculum, in terms of: Pedagogy, Physics Content, and Professional Development. Figure by E. Alicea-Muñoz.

Conclusions

The creation of the CETL 8000 PH course noticeably improved the amount and variety of preparation for new GTAs, in terms of physics content, pedagogy, and continuous feedback, giving the new GTAs more tools to become better teachers and enhance their physics careers.

Moreover, the GTAs who had been trained with the CETL 8000 PH gave feedback about the parts that they liked more or that they found more useful for their teaching, allowing the instructors to adapt the training course to the GTAs' needs and evolving the CETL 8000 PH to a better training course every year.

The parts that the GTAs found the most useful were microteaching, classroom observations and midterm evaluations, which were the parts where they received more feedback (from students, other new GTAs and experienced GTAs). This knowledge about the preferences of the GTAs entailed an evolution to the training program, adding new parts, such as the peer observations, and it will continue to ensure that the CETL 8000 PH covers all the needs of the new GTAs, and therefore, to help them to become better teachers.

Now, we want to discuss if this improvement on their preparation is not only noticed by the GTAs, but reflected in an improvement of their teaching skills, as reported by student evaluations.

Part II
Methodology

3. Analysis

Once concluded the first three years of the CETL 8000 PH, we had already seen that the course was beneficial for the GTAs, as they had reported, and it had evolved towards their needs, trying to improve those topics where the GTAs had more troubles or needed more training.

But now, our main goal is to be sure that this effectiveness on the GTAs is also helpful for their students and make sure that they receive a better teaching from their GTAs.

To do that, we are going to analyze the results of the Teaching Assistant Opinion Survey (TAOS), which is a tool designed to specifically to collect data related to student perceptions of their teaching assistants. This survey has been running since Fall 2011, so we have two years of data before the implementation of CETL 8000 PH and three years of data after it.

One question that could appear is why don't we analyze the students' learning outcomes (exams, assignments, etc.) instead of only studying the surveys. First of all because, although the GTAs are a big influence in their students learning, the students come from a wide variety of backgrounds (e.g., major and incoming GPA) and they also attend lecture courses taught by faculty, so it will be extremely difficult to isolate the effect of GTAs' performance on the students' grades [9].

Moreover, even if the students were only taught by the GTAs, there is no clear evidence of a relation between their grades and the evaluation of their teachers, probably due to the big amount of variables that could influence the final grade of the students [3].

4. Data collection (TAOS survey)

All data used in this analysis was collected from the TAOS. The survey has 12 questions:

- **Approachability:** The GTA was accessible for assistance during the course (Strongly Disagree → Strongly Agree).
- **Attitude about teaching:** The GTA's attitude about their teaching role in this course (Detached → Extremely Enthusiastic).
- **Classroom management:** The GTA's management of classroom/lab environment (Very Poor → Exceptional).
- **Concept familiarity:** The GTA's familiarity with course concepts (Very Poor → Exceptional).
- **Engaged students:** The GTA actively engaged students, for example with questions, participation, group work, etc. (Very Poor → Exceptional).
- **Explained concepts clearly:** The GTA explained course concepts clearly (Strongly Disagree → Strongly Agree).
- **Oral communication:** The GTA's oral communication skills (Very Poor → Exceptional).
- **Overall effectiveness:** Considering everything, the GTA was an effective GTA (Strongly Disagree → Strongly Agree).
- **Preparedness:** The GTA's level of preparation (Completely Unprepared → Extremely Well Prepared).
- **Respect for students:** The GTA's respect for their students (Very Poor → Exceptional).
- **Stimulated interest:** The GTA stimulated my interest in the subject matter (Ruined My Interest → Made Me Eager to Learn More).
- **Written communication:** The GTA's written communication skills (Very Poor → Exceptional).

Each question is rated in a 5 points scale ($1 \rightarrow 5$) and the final rate for every question is the interpolated median of all the results.

The TAOS survey was designed to give feedback to the Teaching Assistants and to the Institute; so, it was not specifically done to analyze the effect of a training program on the GTAs, but the questions of the TAOS can evaluate significantly the performance of the GTAs and, therefore, we can extract valuable information of them. Another important point about them is that the questions have remained the same since 2011, so we can analyze, for every question, the evolution among the five years of results.

5. Statistical analysis methods

In order to analyze the data from the TAOS survey, we will need several statistical methods, keeping in mind that our main goal is to determine if there are statistically significant differences between the means, of the first year of teaching, of the GTAs who were trained using the CETL 8000 PH (Post group, starting Fall 2013, 2014 and 2015), and those who didn't (Pre group, starting Fall 2011 and 2012). We also want to study the relations between GTAs who started teaching in different semesters (from Fall 2011 to 2015) and see if there are significant differences on their TAOS scores.

In addition, we will also consider other possible effects on their scores such as their nationality (if they are from the United States or not) or their teaching assignment (Traditional vs M&I).

Moreover, we will also study the evolution of the GTAs among all their semesters of teaching, and compare those who took the CETL 8000 PH and those who didn't.

Data

After the 5 years of data collecting, we have a total of 336 scores for every TAOS question; 233 of these are for first-time GTAs, that is where we are going to do more emphasis on our analysis. These data are equally distributed between the pre-post groups, national-international GTAs and Traditional-M&I course, as the Table 1 shows. The data are not as well distributed among the 5 different starting years of teaching but still we have enough to perform some analysis as we can see in the Table 2. We can also see (Table 3) that the national and international GTAs are equally distributed between the Traditional and the M&I, so in principle, a difference in one subgroup should not affect the other.

Total	Pre	Post	**National**	Pre	Post	**International**	Pre	Post
First Fall	51	69	First Fall	25	45	First Fall	21	24
First Spring	49	64	First Spring	23	38	First Spring	22	26
			Traditional	Pre	Post	**M&I**	Pre	Post
			First Fall	15	29	First Fall	29	40
			First Spring	18	26	First Spring	25	38

Table 1: Distribution of the GTAs on their first year of teaching, divided by fall-spring semesters, pre-post groups, nationality and Traditional-M&I.

It is important to highlight that the analysis will be performed by semesters (and not by years) so there will not be repeated GTAs in each group, which would lead to a non independence of the data inside the groups.

Shapiro-Wilk

One of the most common assumptions in many statistical tests is the normality of the data, and thus, we need a test to check for this assumption in our case. The Shapiro-Wilk test is used to determine if a given distribution is normal or not and, according to [13], it is the most powerful test for this purpose. As many other similar tests, it doesn't work as well with small samples, but it is still a good test for them [14].

This test is based in the null hypothesis that the data provided does correspond to a normal distribution, and if it is statistically significant we reject that hypothesis.

Total	2011	2012	2013	2014	2015
First Fall	16	35	20	13	36
First Spring	18	31	18	14	32
National	2011	2012	2013	2014	2015
First Fall	8	17	16	6	23
First Spring	8	15	14	6	18
International	2011	2012	2013	2014	2015
First Fall	7	14	4	7	13
First Spring	8	14	4	7	14
Traditional	2011	2012	2013	2014	2015
First Fall	5	10	11	2	16
First Spring	7	11	10	2	14
M&I	2011	2012	2013	2014	2015
First Fall	10	19	9	11	20
First Spring	9	16	8	12	18

Table 2: Distribution of the GTAs on their first year of teaching, divided by fall-spring semesters, starting semester of teaching, type of course and nationality.

First Fall	**Pre**	National	Intarnational	**Post**	National	Intarnational
	Traditional	7	8	Traditional	19	10
	M&I	17	12	M&I	26	14
First Spring	**Pre**	National	Intarnational	**Post**	National	Intarnational
	Traditional	8	10	Traditional	18	8
	M&I	14	11	M&I	20	18

Table 3: Distribution of the GTAs on their first year of teaching, divided by fall-spring semesters, pre-post groups, national-international and Traditional-m&I.

The test is statistically significant if the statistic W $<$ p-value, where this p-value is fixed arbitrarily. In this case, and for all the analysis that we will perform, the p-value will be: p-value $= 0.05$, which is the usual value given in the majority of tests.

Mann-Whitney

The Mann-Whitney test in a non parametric test used to determine if two distributions are equal or not. It is based on the null hypothesis that the distributions of two samples are the same, and thus, if it is statistically significant we reject the null hypothesis in favor of the alternative hypothesis, which is that they are different [11]. The main difference with other tests, like the independent samples t-test, is that the distributions of the two samples don't have to be normally distributed, which will be very useful in our cases. If the distributions of the two groups are the same, with this test we will be able to determine if there are statistically significant differences between their means [1]. To do this test we need to check previously some assumptions, ones related to the study and other to the data. The assumptions about the study are that we only have one dependent variable and one independent variable which is consists of two different groups. In our case the dependent variable will always be the TAOS score for the different categories of the survey; and the independent variable will be the Pre and Post groups, the nationality of the students (from the United States or not) and the course type (Traditional or M & I). As the different groups are independent, that is to say that there aren't students in both groups at the same time, we have independence of observations, which is also an assumption for this test. The assumption related to the data is that the distribution of both groups have to have the similar shapes (which doesn't have to be normal) in order to perform a good analysis. If this assumption is violated, the Mann-Whitney test can still be used to compare mean ranks.

One-Way ANOVA

The one-way analysis of variance (one-way ANOVA) is a test used to compare the means of three or more independent groups. It can also be used with only two groups, but in this case the result is the same that with an independent samples t-test. This test is based on the F statistic and its null hypothesis is that the means of the different groups are the same; so if it is statistically significant it means that at least one group is statistically significant to another, but the test does not say which one. To determine that we need to do a post hoc analysis. To perform this test, the data has to fit some assumptions. The dependent variable has to be continuous and the independent variable should consist of three or more independent groups. We also need independence of observations, which in our case it is fulfilled as the groups are the different starting semesters of the GTAs and, obviously, it can't be the same GTA in more than one group. The other assumptions are related to the data itself. There cannot be significant outliers (extreme values of the data), which could interfere in the results. The variances of the different groups have to be equal (or similar). This assumption is easily checked with the Levene's test, which we will explain in this section, and in the case that it is not fulfilled (we have heterogeneity of variances) we can perform a Welch test, instead of the ANOVA, to obtain the results. Finally, we need the dependent variable to be approximately normally distributed for every group of the independent variable. We will use the Shapiro-Wilk test to check this assumption, but considering that the ANOVA is pretty robust to violations of the normality.

Levene test

The Levene test is a statistical test used to determine if two or more samples have the same variance. It is based in the null hypothesis that the variances are equal among all the independent groups, and if it is significant we reject this hypothesis in favor of the alternative hypothesis that is that at least two samples are not likely to have the same variance [10] We will use this test to determine if our data has homogeneity of variances in order to decide which method should we use; as in some tests it is required in the interest of having correct results.

Welch

The Welch test is an adaption of the students t-test in the case that the samples have different variances and different sizes [16], so is a test with the null hypothesis that two samples have the same means, but in this case the assumption of homogeneity of variances does not have to be fulfilled. If the test is significant we reject the null hypothesis and accept the alternative which is that the two groups have different means. This test can be extended to the case where we have more than two independent groups [17], being then an adaptation of the one-way ANOVA in the cases when the assumption of equality of variances is violated. We will use this test in that last case, as some of the data will not have homogeneity of variances when performing the one-way ANOVA.

Kruskal-Wallis

The Kruskal-Wallis test is a non-parametric test used to determine if different samples came from the same distribution [8]. It can be used with more than two groups and it is not required normality of the samples; so it could be interpreted like an extension of the Mann-Whitney test for more than two groups or an extension of the one-way ANOVA for non normally distributed samples. As the one-way ANOVA, the null hypothesis is that all the groups have the same means, and when it is rejected it means that al least we have two groups with different means, but without knowing which of them. We should perform a post-hoc analysis in order to know which groups have different means. The assumptions needed for this test are more or less the same that for the Mann-Whitney: we need one dependent variable, one independent variable which consists of two or more different groups, and independence of observations. The distributions of the different groups of the independent variable have to have similar shapes in order to compare the means of the different groups.

Two-Way ANOVA

The two-way analysis of variances (two-way ANOVA) is an extension of the one-way ANOVA for when we have two independent variables, instead of a single one. This test compares the means among the different groups of each independent variable, and determines if they have the same mean or not. It also determines the relation between the two independent variables. The assumptions to this test are the same as the one-way ANOVA. We will use this test to compare the effects of the different independent variables (pre-post groups, national-international students and Traditional-M&I course) studied individually before with the other tests. Our goal with this test is determine which effects are bigger and look for relations between them.

Part III
Results

In this section we will be discussing the results of the analysis. First of all, we will analyze exhaustively the effect of the CETL 8000 PH between the GTAs who didn't take that preparation program (Pre group: starting Fall 2011 and 2012) and the GTAs who did take the training program (Post group: starting Fall 2013, 2014 and 2015). In the first part of the analysis, we will be only looking to the first year scores of the GTAs, divided by first fall and first spring to avoid dependences between them. That's because is where the training program should have more effect and it's that first year where we have more data to analyze, as the GTAs are usually assigned to other courses after that first year. We will focus on the effect on different subgroups such as the nationality (distinguishing between GTAs from the United States and GTAs from other countries) and the course type (Traditional and M&I) to see in which ones the CETL 8000 PH is more effective. We will also discuss if there are statistically significant differences between the scores of the GTAs from different subgroups, without taking into account if the GTAs have gone trough the training program or not.

After that, we will look to the effect of the training program according to the first year of teaching of the GTAs. The main objective with this test will be to see if there are differences between the 3 different cycles of the CETL 8000 PH.

We will also analyze the effect of the different independent variables (pre-post groups, nationality) on the TAOS scores, and determine which ones are more important.

Finally, we will analyze the evolution of the GTAs' scores across the different semesters of teaching. This is the only part where we will be utilizing the non-first year data, and we won't be doing the same tests as before because of the impossibility of performing the same test with this data.

All the test will be performed for every question of the TAOS survey and then, we will have 12 different categories to analyze each time. The main objective of the whole analysis is not only to see if the CETL 8000 is an efficient training program, but see in which aspects it is better or worst and see the different effects on different aspects for every subgroup of GTAs.

All these results will be presented on the 2017 Winter Meeting of the American Association of Physics Teachers, in Atlanta, GA, on 20 February 2017.

6. First Year one independent variable

In this section we will study only the results of the first year of teaching of the GTAs, considering only one independent variable for each case.

6.1 Normality tests

One of the most common assumptions for many test is the normality of the data, and thus, we need to know how it is our data in order to perform correctly the different tests.

Figures 3 and 4 show the histograms of the different TAOS questions for the fall and spring semester, and Tables 4 and 5 show the Shapiro-Wilk results for the total, pre and post data of the fall and spring semesters.

As we can see in the histograms and in the results of the Shapiro-Wilk test, our data is clearly not normal for any of the different categories (except for the stimulated interest of the pre group) so we must use non-parametric tests in order to have consistent results. The majority of the distributions have higher Kurtosis coefficients in the spring semester and all of them are skewed to the right. This behavior about the Kurtosis coefficient could explain why the spring semester results are usually more significant than the results of the fall semester, as we will see in the following sections.

	Skewness	Kurtosis	Total p-value	**Pre p-value**	**Post p-value**
Approachability	-1.670	3.246	0.001	0.001	0.001
Attitude about teaching	-0.897	0.764	0.001	0.003	0.001
Classroom management	-1.373	2.310	0.001	0.001	0.001
Concept familiarity	-1.903	4.093	0.001	0.001	0.001
Engaged students	-0.096	0.607	0.001	0.011	0.001
Explained concepts clearly	-1.015	0.832	0.001	0.002	0.001
Oral communication	-0.688	-0.371	0.001	0.005	0.001
Overall effectiveness	-1.065	0.812	0.001	0.001	0.001
Preparedness	-0.185	5.076	0.001	0.001	0.001
Respect for students	-1.781	3.537	0.001	0.001	0.001
Stimulated interest	-0.410	-0.127	0.015	0.446	0.071
Written communication	-0.949	1.175	0.001	0.059	0.001

Table 4: Shapiro-Wilk normality test results for the fall semester data. Skewness, kurtosis and p-values for the total data and p-values for the pre and post groups.

	Skewness	Kurtosis	Total p-value	**Pre p-value**	**Post p-value**
Approachability	-1.979	5.500	0.001	0.001	0.001
Attitude about teaching	-2.401	10.143	0.001	0.002	0.001
Classroom management	-1.716	4.085	0.001	0.008	0.001
Concept familiarity	-1.364	1.677	0.001	0.007	0.001
Engaged students	-1.774	4.651	0.001	0.001	0.001
Explained concepts clearly	-1.383	2.770	0.001	0.002	0.001
Oral communication	-1.047	1.229	0.001	0.017	0.001
Overall effectiveness	-2.603	10.004	0.001	0.001	0.001
Preparedness	-1.882	5.629	0.001	0.001	0.001
Respect for students	-1.946	5.201	0.001	0.001	0.001
Stimulated interest	-0.537	0.136	0.002	0.364	0.001
Written communication	-0.983	1.018	0.001	0.093	0.001

Table 5: Shapiro-Wilk normality test results for the spring semester data. Skewness, kurtosis and p-values for the total data and p-values for the pre and post groups.

6.2 Comparing different independent variables

Before to start analyzing in depth the effect of the CETL 8000 PH, we want to see if there are significant differences between the scores of the national and international GTAs and between the Traditional and M&I course. If there weren't any differences, it wouldn't be necessary to look the effect of the training

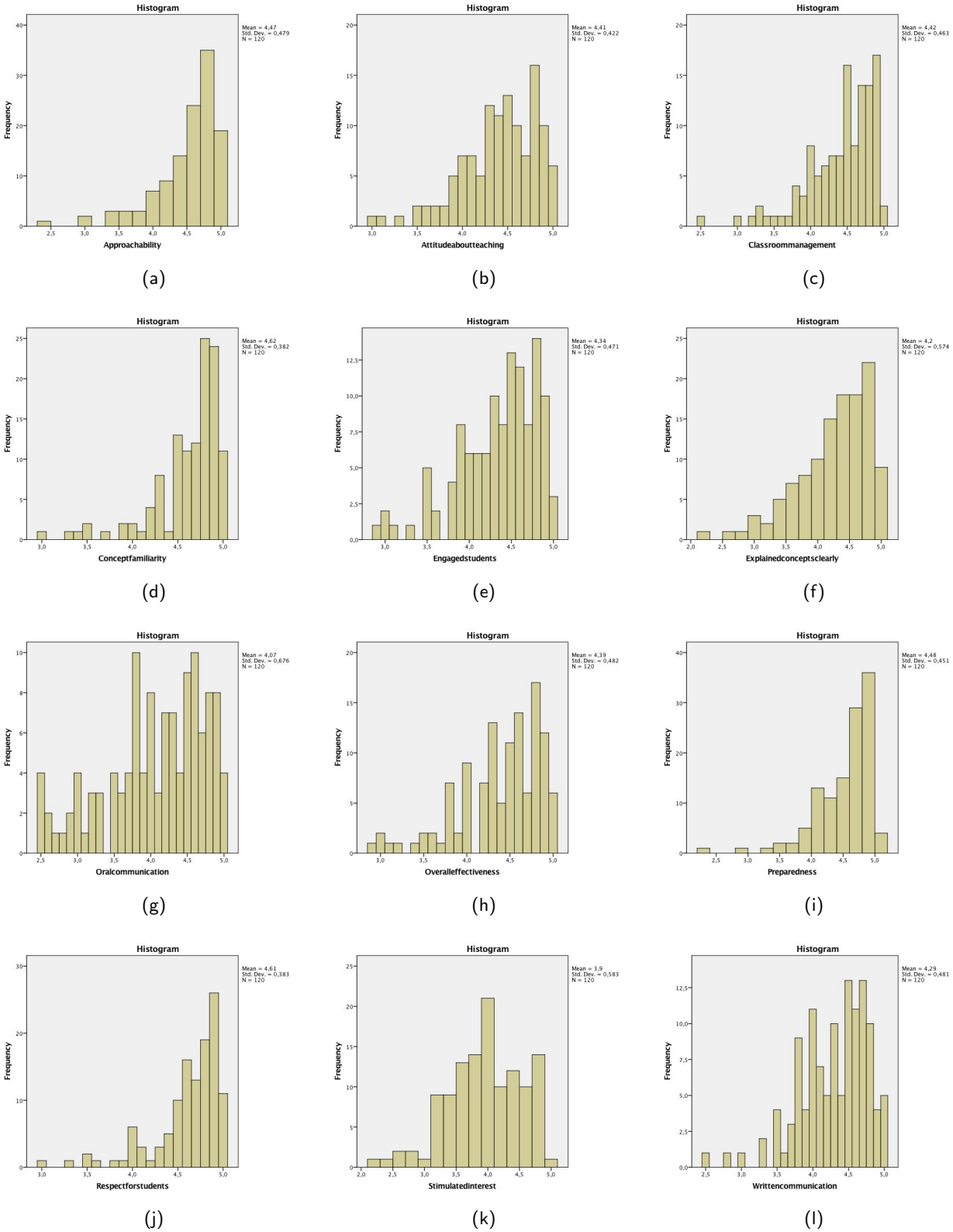


Figure 3: Histograms of the distributions for all the sections in the fall semester.

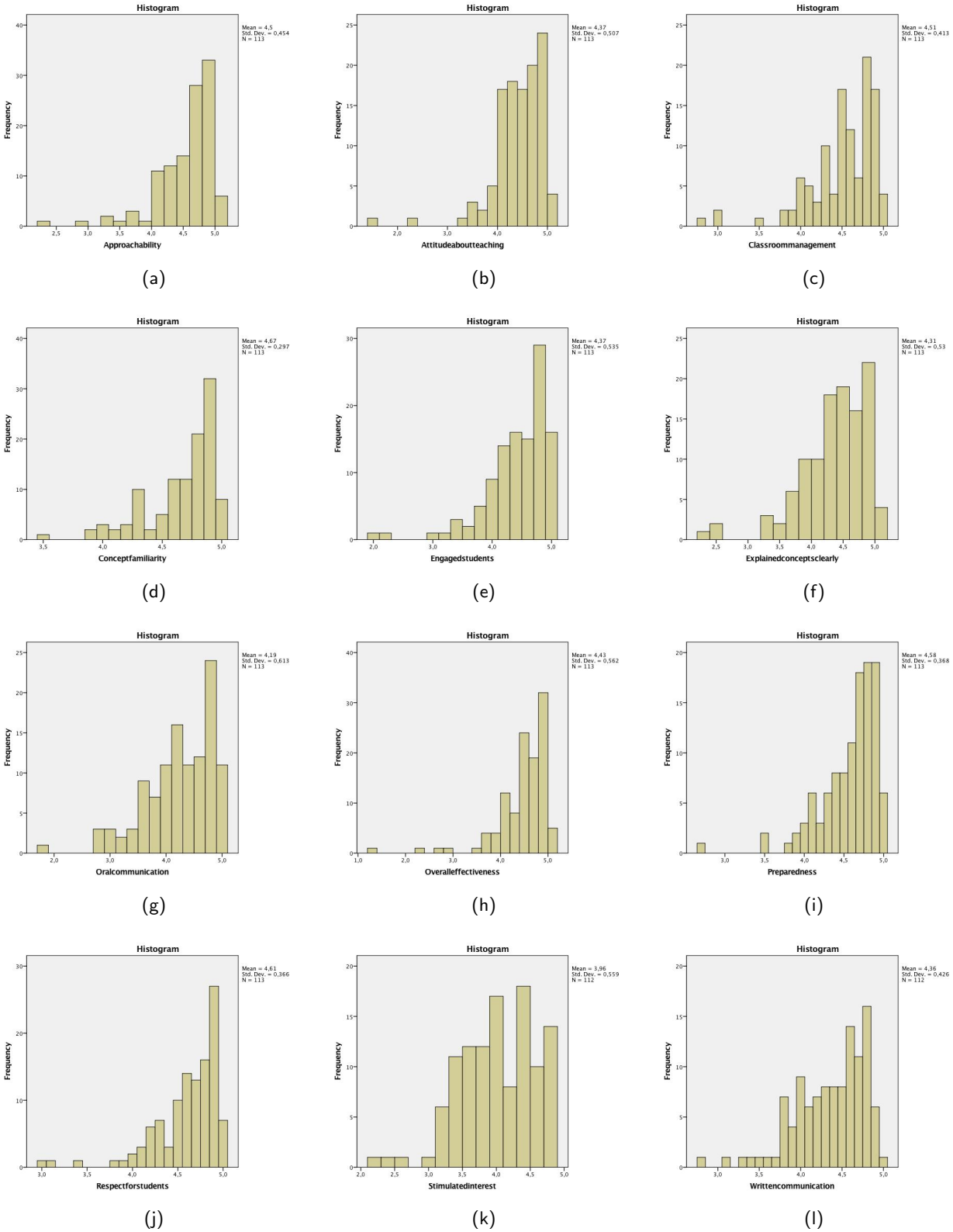


Figure 4: Histograms of the distributions for all the sections in the spring semester.

program in each subgroup. To do that we will use the Mann-Whitney test with the nationality and the type of course as the independent variables and the TAOS scores for each category as dependent variable (this will be the dependent variable through all the tests). The only assumption that we need to check in order to perform the test is the equality of the distributions, so in this case the Mann-Whitney test would be a test for the difference of means. This proof is done in the section A.1 of the Appendix.

	National	Interational	**Fall**	National	Interational	**Spring**
Approachability	4.571	4.344	0.008	4.625	4.413	0.001
Attitude about teaching	4.499	4.311	0.044	4.439	4.333	0.066
Classroom management	4.543	4.276	0.002	4.585	4.454	0.004
Concept familiarity	4.716	4.498	0.002	4.746	4.590	0.001
Engaged students	4.489	4.127	0.001	4.518	4.250	0.001
Explained concepts clearly	4.460	3.833	0.001	4.513	4.110	0.001
Oral communication	4.461	3.502	0.001	4.516	3.833	0.001
Overall effectiveness	4.556	4.173	0.001	4.580	4.310	0.001
Preparedness	4.589	4.356	0.035	4.641	4.546	0.031
Respect for students	4.661	4.564	0.059	4.670	4.579	0.007
Stimulated interest	4.107	3.642	0.001	4.146	3.787	0.001
Written communication	4.473	4.036	0.001	4.510	4.204	0.001

Table 6: Mann-Whitney test for the first fall and spring smesters. Means of the National and International groups, and p-value of the Mann-Whitney test.

	Traditional	M&I	**Fall**	Traditional	M&I	**Spring**
Approachability	4.609	4.419	0.001	4.614	4.479	0.124
Attitude about teaching	4.450	4.414	0.364	4.430	4.373	0.819
Classroom management	4.570	4.367	0.001	4.641	4.452	0.011
Concept familiarity	4.730	4.584	0.004	4.727	4.648	0.139
Engaged students	4.398	4.314	0.104	4.445	4.381	0.256
Explained concepts clearly	4.227	4.219	0.394	4.332	4.340	0.760
Oral communication	4.098	4.088	0.665	4.218	4.219	0.384
Overall effectiveness	4.466	4.372	0.064	4.486	4.452	0.829
Preparedness	4.616	4.443	0.002	4.652	4.568	0.302
Respect for students	4.709	4.565	0.005	4.700	4.581	0.076
Stimulated interest	3.950	3.916	0.680	3.993	3.989	0.928
Written communication	4.350	4.296	0.349	4.400	4.365	0.668

Table 7: Mann-Whitney test for the first fall and spring smesters. Means of the Traditional and M&I groups, and p-value of the Mann-Whitney test.

The results of the test are shown in the Tables 6 and 7 and we can see how in the case of the National and International groups we have statistically significance for all the categories except the Respect for students of the fall semester and the Attitude about teaching of the spring semester. On the other case we can see that, although there are not that many categories with a statistically significant difference (especially in the spring semester) between the Traditional and the M&I course, we can also see that there are some differences between them and thus, it makes sense to look the effect of the CETL 8000 PH in each one if the courses separately.

6.3 Mann-Whitney test by pre and post groups

In this section we want to discuss if there are statistically significant differences between the pre (starting Fall 2011 and 2012) and post (starting Fall 2013, 2014 and 2015) groups.

First of all, we will analyze all the results (total), and after that we will analyze the results dividing by national-international students and Traditional-M&I groups. We have seen that there are statistically significant differences between these two subgroups and now, we want to see if the training program has different effect in one or another subgroup, and if so, which categories are more affected by the training program.

The only assumption that we have to check is that the distributions of the pre and post group, for each question and for every subgroup, and have similar shapes. This is done in the section A.2 of the Appendix.

The other assumptions for the Mann-Whitney test are fulfilled as we have one dependent variable (the scores) and only two independent groups, which will be different in every case. The data inside every group is also independent as the GTAs are divided in fall and spring semesters and we are only analyzing their first year of teaching.

In the following tables 8, 9, 10. 11 and 12 we can see the means of the pre and post groups and p-values of the Mann-Whitney test. The test will be statistically significant if the p-value <0.05.

We can see in the table 8 that we have statistically significant differences in almost every question of the fall semester and in all of the spring semesters, and in all the cases the mean of the post group is higher than the mean of the pre group. So we can see an important improvement in all the questions.

In the National 9 and International 10 tables we can see that the significances are not the same; we only have p-values <0.05 in one question of the fall semester in both national and international, while in the spring semester we have much more questions with statistically significant differences. This differences between the p-values of the fall and spring semesters are repeated along all the analysis. They are due to a slightly diminishment in the scores of the pre group and and improvement in the scores of the post group, from the fall to the spring semester. We can also see that the p-values of the test are, in general, higher in these two cases because we have less data in each case; as the means differences are similar to the total case (except in some specific questions).

In the other analysis, dividing between Traditional 11 and M&I 12 courses, we can also see that the p-values are consistently lower in the spring semester. A part from this, the important conclusions we can obtain is that there are some questions in which the training program has more effect in one course that the other; these are for example the oral communication and the explained concepts clearly which the p-values are clearly smaller for the Traditional course, and thus, the CETL 8000 PH is more effective, in this two aspects, on the Traditional course.

In all the cases of this analysis, the means of the post group are higher than the pre ones, so in all analysis a smaller p-value means more difference between the means, and a higher effect of the training program.

6.4 Analysis by semesters

In this section we will focus in the differences between semesters, instead of pre and post groups as we have done before. We will only perform the analysis with all the GTAs who teach for first time, without distinguishing between national-international and Traditional-M&I courses because, as we can see on the table 2, there are some semesters where we don't have enough data to have consistent results.

TOTAL	**Pre**	**Post**	**Fall p-value**	**Pre**	**Post**	**Spring p-value**
Approachability	4.341	4.559	0.014	4.410	4.577	0.010
Attitude about teaching	4.320	4.474	0.051	4.235	4.477	0.001
Classroom management	4.322	4.488	0.103	4.392	4.597	0.001
Concept familiarity	4.482	4.716	0.003	4.576	4.741	0.001
Engaged students	4.235	4.419	0.027	4.253	4.464	0.007
Explained concepts clearly	4.051	4.314	0.022	4.135	4.445	0.001
Oral communication	3.908	4.184	0.030	4.059	4.298	0.027
Overall effectiveness	4.269	4.478	0.029	4.292	4.544	0.001
Preparedness	4.353	4.574	0.027	4.459	4.664	0.001
Respect for students	4.525	4.674	0.053	4.492	4.695	0.001
Stimulated interest	3.745	4.017	0.017	3.783	4.095	0.002
Written communication	4.163	4.386	0.013	4.210	4.475	0.001

Table 8: Mann-Witney test results for the fall and spring semesters. Means of the Pre and Post groups, and p-value of the Mann-Whitney test.

NATIONAL	**Pre**	**Post**	**Fall p-value**	**Pre**	**Post**	**Spring p-value**
Approachability	4.512	4.604	0.156	4.596	4.642	0.164
Attitude about teaching	4.460	4.520	0.409	4.357	4.489	0.035
Classroom management	4.484	4.576	0.381	4.470	4.655	0.009
Concept familiarity	4.600	4.780	0.025	4.674	4.789	0.016
Engaged students	4.404	4.536	0.154	4.435	4.568	0.026
Explained concepts clearly	4.408	4.489	0.200	4.383	4.592	0.005
Oral communication	4.452	4.467	0.453	4.474	4.542	0.100
Overall effectiveness	4.512	4.580	0.405	4.535	4.608	0.032
Preparedness	4.548	4.611	0.266	4.552	4.695	0.049
Respect for students	4.588	4.702	0.317	4.574	4.729	0.006
Stimulated interest	4.020	4.156	0.271	4.004	4.232	0.022
Written communication	4.424	4.500	0.204	4.417	4.566	0.018

Table 9: Mann-Witney test results for the national GTAs of the fall and spring semesters. Means of the Pre and Post groups, and p-value of the Mann-Whitney test.

INTERNATIONAL	Pre	Post	Fall p-value	Pre	Post	Spring p-value
Approachability	4.195	4.475	0.138	4.332	4.481	0.151
Attitude about teaching	4.224	4.388	0.193	4.186	4.458	0.034
Classroom management	4.219	4.325	0.584	4.386	4.512	0.200
Concept familiarity	4.386	4.596	0.165	4.495	4.669	0.041
Engaged students	4.043	4.200	0.226	4.177	4.312	0.324
Explained concepts clearly	3.657	3.988	0.101	3.968	4.231	0.055
Oral communication	3.329	3.654	0.091	3.705	3.942	0.106
Overall effectiveness	4.043	4.287	0.118	4.145	4.450	0.020
Preparedness	4.186	4.504	0.104	4.459	4.619	0.047
Respect for students	4.500	4.621	0.220	4.500	4.646	0.040
Stimulated interest	3.510	3.758	0.108	3.652	3.896	0.122
Written communication	3.881	4.171	0.034	4.033	4.342	0.004

Table 10: Mann-Witney test results for the international GTAs of the fall and spring semesters. Means of the Pre and Post groups, and p-value of the Mann-Whitney test.

TRADITIONAL	Pre	Post	Fall p-value	Pre	Post	Spring p-value
Approachability	4.507	4.662	0.310	4.494	4.696	0.023
Attitude about teaching	4.293	4.531	0.049	4.261	4.546	0.027
Classroom management	4.533	4.590	0.425	4.550	4.704	0.035
Concept familiarity	4.720	4.734	0.464	4.611	4.808	0.004
Engaged students	4.227	4.486	0.038	4.311	4.538	0.050
Explained concepts clearly	3.920	4.386	0.031	4.106	4.488	0.004
Oral communication	3.753	4.276	0.021	4.000	4.369	0.035
Overall effectiveness	4.333	4.534	0.119	4.300	4.615	0.009
Preparedness	4.593	4.628	0.587	4.578	4.704	0.084
Respect for students	4.593	4.769	0.078	4.578	4.785	0.007
Stimulated interest	3.740	4.059	0.099	3.772	4.146	0.040
Written communication	4.200	4.428	0.104	4.239	4.512	0.011

Table 11: Mann-Witney test results for the Traditional GTAs of the fall and spring semesters. Means of the Pre and Post groups, and p-value of the Mann-Whitney test.

M&I	Pre	Post	Fall p-value	Pre	Post	Spring p-value
Approachability	4.328	4.485	0.093	4.456	4.495	0.299
Attitude about teaching	4.390	4.432	0.696	4.288	4.429	0.034
Classroom management	4.300	4.415	0.417	4.344	4.524	0.012
Concept familiarity	4.421	4.703	0.003	4.576	4.695	0.050
Engaged students	4.238	4.370	0.272	4.332	4.413	0.133
Explained concepts clearly	4.159	4.263	0.669	4.224	4.416	0.039
Oral communication	4.048	4.118	0.865	4.172	4.250	0.394
Overall effectiveness	4.283	4.437	0.282	4.388	4.495	0.044
Preparedness	4.317	4.535	0.024	4.464	4.637	0.017
Respect for students	4.510	4.605	0.520	4.500	4.634	0.007
Stimulated interest	3.817	3.987	0.262	3.875	4.061	0.077
Written communication	4.214	4.355	0.222	4.229	4.450	0.008

Table 12: Mann-Witney test results for the M&I GTAs of the fall and spring semesters. Means of the Pre and Post groups, and p-value of the Mann-Whitney test.

We have already seen that there is a significant improvement between the pre (starting Fall 2011 and 2012) and post (starting 2013, 2014 and 2015) groups, and now we want to focus on the differences inside these groups, to see if the different cycles of the CETL 8000 PH have different results on the TAOS scores.

To do that we will perform a ANOVA test, with the Welch correction when needed (when the Levene test for equally of variances is significant), and a Kruskal-Wallis test, as our data is not normally distributed. We will perform these tests between the 2013, 2014 and 2015 years.

We can see in the tables 13 and 14 that the test results are very similar between the ANOVA, Welch and Kruskal-Wallis, even though our distributions are not normal. This results could be due to the fact that we don't have to much data, and also to the fact that the one-way ANOVA is considerably robust to normality violations.

All the results, except the stimulated interest of the spring semester, show that there are not statistically significant differences between the three different groups. If we look to the means of the different groups, we can see how in the fall semester the 2014 groups means are slightly smaller that the other two, while on the other hand, in the spring semester the 2014 means are slightly higher that the other two. We will discuss more exhaustively these facts on the discussion section and we will try to explain why we have them; but in a first approach, we can say that the three cycles of CETL 8000 PH have similar impact on the GTAs teaching skills, as perceived by their students according to the results of the TAOS survey.

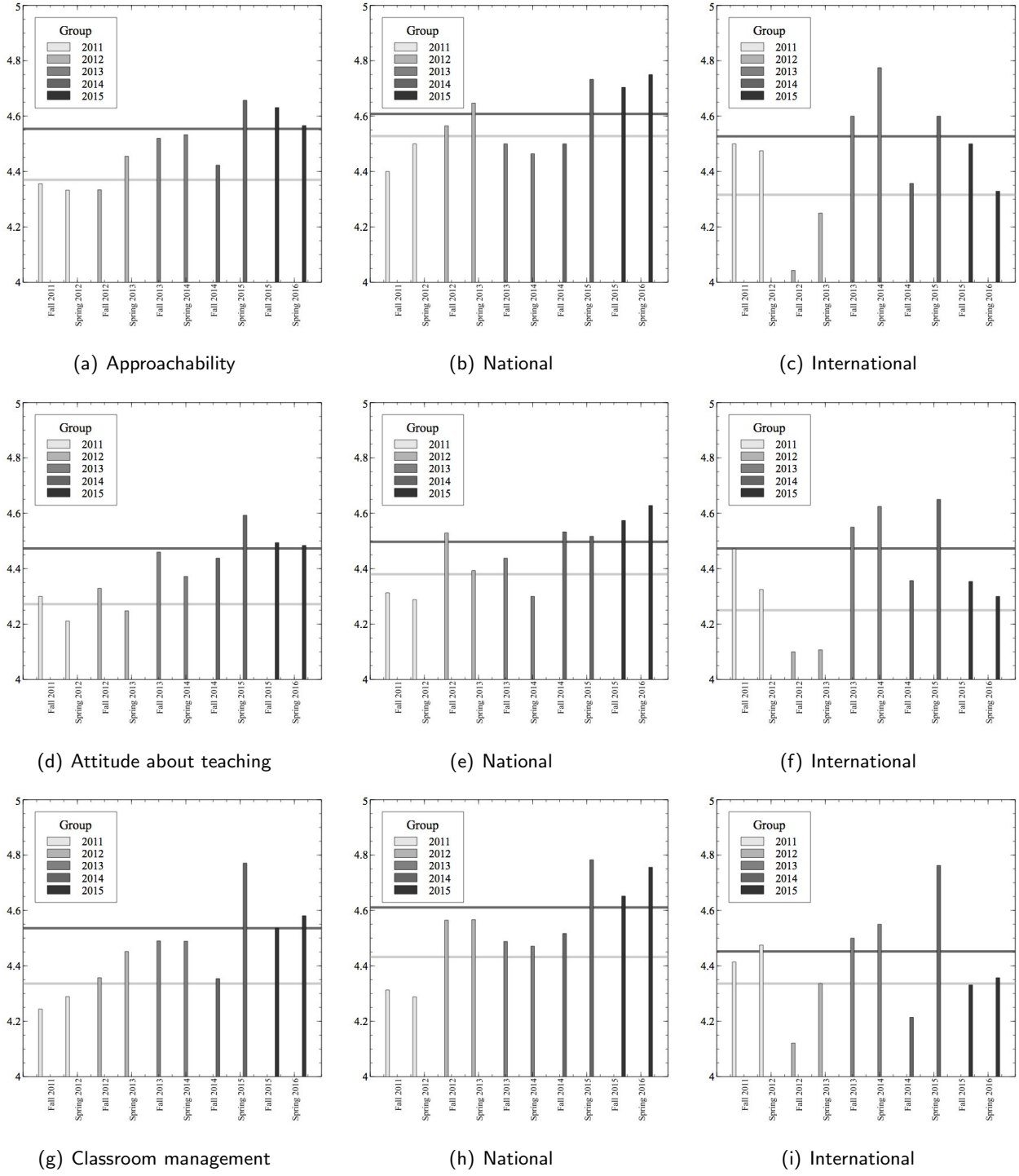
In the Figures 5, 6, 7 and 8 we can see the means of the fall and spring semester of the different starting semester groups. There is shown in first place the means of all the GTAs and then there is a plot for the national and the international GTAs. The horizontal lines are the total means (fall and spring semesters) of the pre (light line) and post (dark line) groups. We can see the evolution of the means and, more importantly, the fact that the mean of the post group is always bigger than the mean of the pre group. We can also observe that the national GTAs have always better scores than the international GTAs.

Fall POST	**2013**	**2014**	**2015**	**ANOVA**	**Levene**	**Welch**	**Kruskal-Wallis**
Approachability	4.520	4.423	4.631	0.239	0.215	0.276	0.197
Attitude about teaching	4.460	4.438	4.494	0.898	0.942	0.904	0.867
Classroom management	4.490	4.354	4.536	0.358	0.747	0.445	0.324
Concept familiarity	4.790	4.669	4.692	0.349	0.056	0.178	0.477
Engaged students	4.435	4.331	4.442	0.735	0.691	0.795	0.739
Explained concepts clearly	4.380	4.238	4.306	0.723	0.243	0.689	0.728
Oral communication	4.325	4.069	4.147	0.466	0.077	0.352	0.533
Overall effectiveness	4.525	4.392	4.483	0.679	0.466	0.728	0.801
Preparedness	4.635	4.600	4.531	0.498	0.260	0.465	0.648
Respect for students	4.645	4.577	4.725	0.331	0.168	0.413	0.645
Stimulated interest	3.965	3.831	4.114	0.239	0.967	0.275	0.239
Written communication	4.390	4.400	4.378	0.986	0.377	0.986	0.922

Table 13: Test results of the fall semesters of the post group, with the starting semester of teaching as de dependent variable. Means of each starting semester group, p-values of the one-way ANOVA, Levene, Welch and Kruskal-Wallis tests.

Spring POST	**2013**	**2014**	**2015**	**ANOVA**	**Levene**	**Welch**	**Kruskal-Wallis**
Approachability	4.533	4.657	4.566	0.726	0.199	0.587	0.732
Attitude about teaching	4.372	4.593	4.484	0.462	0.095	0.380	0.553
Classroom management	4.489	4.771	4.581	0.133	0.063	0.016	0.095
Concept familiarity	4.739	4.800	4.716	0.641	0.206	0.371	0.505
Engaged students	4.361	4.636	4.447	0.324	0.090	0.135	0.498
Explained concepts clearly	4.417	4.636	4.378	0.245	0.117	0.048	0.336
Oral communication	4.267	4.464	4.244	0.464	0.052	0.223	0.818
Overall effectiveness	4.406	4.700	4.553	0.344	0.041	0.132	0.639
Preparedness	4.572	4.793	4.659	0.142	0.077	0.044	0.161
Respect for students	4.611	4.729	4.728	0.488	0.129	0.680	0.962
Stimulated interest	3.956	4.407	4.038	0.043	0.133	0.008	0.043
Written communication	4.394	4.643	4.447	0.213	0.130	0.051	0.259

Table 14: Test results of the spring semesters of the post group, with the starting semester of teaching as de dependent variable. Means of each starting semester group, p-values of the one-way ANOVA, Levene, Welch and Kruskal-Wallis tests.

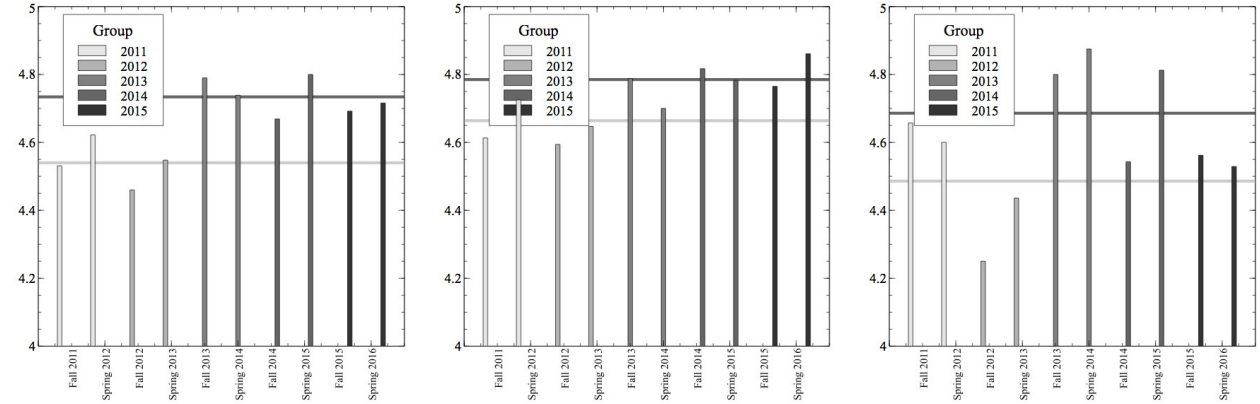


(a) Approachability (b) National (c) International

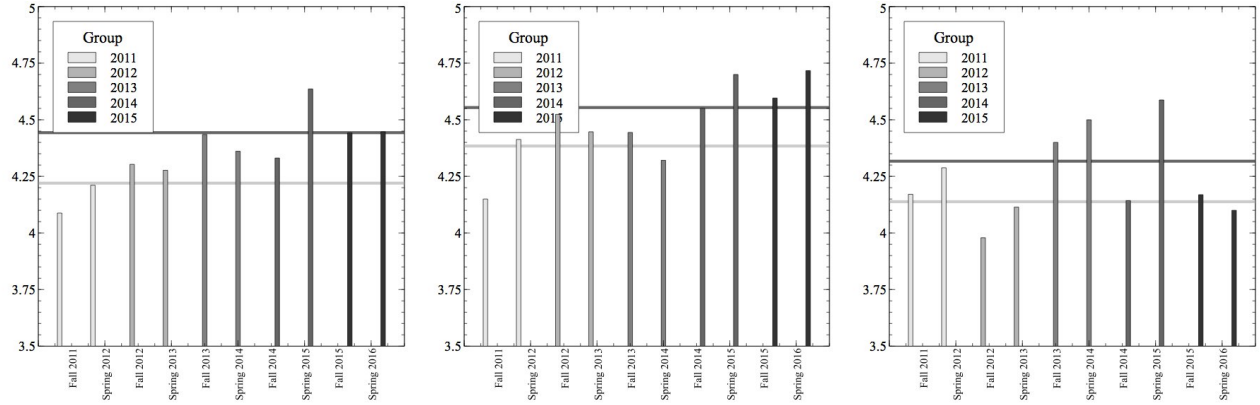
(d) Attitude about teaching (e) National (f) International

(g) Classroom management (h) National (i) International

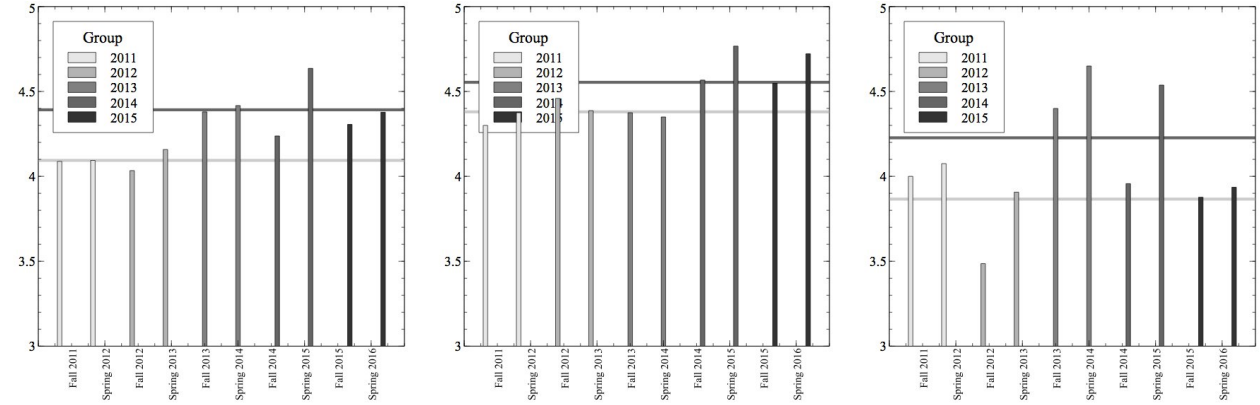
Figure 5: Each figure shows the means for every starting semester group, in their first fall and spring semester of teaching. The light horizontal line is the mean for the pre group(starting Fall 2011 and 2012) and the dark horizontal line is the mean for the post group (starting Fall 2013, 2014 and 2015).



(a) Concept familiarity (b) National (c) International

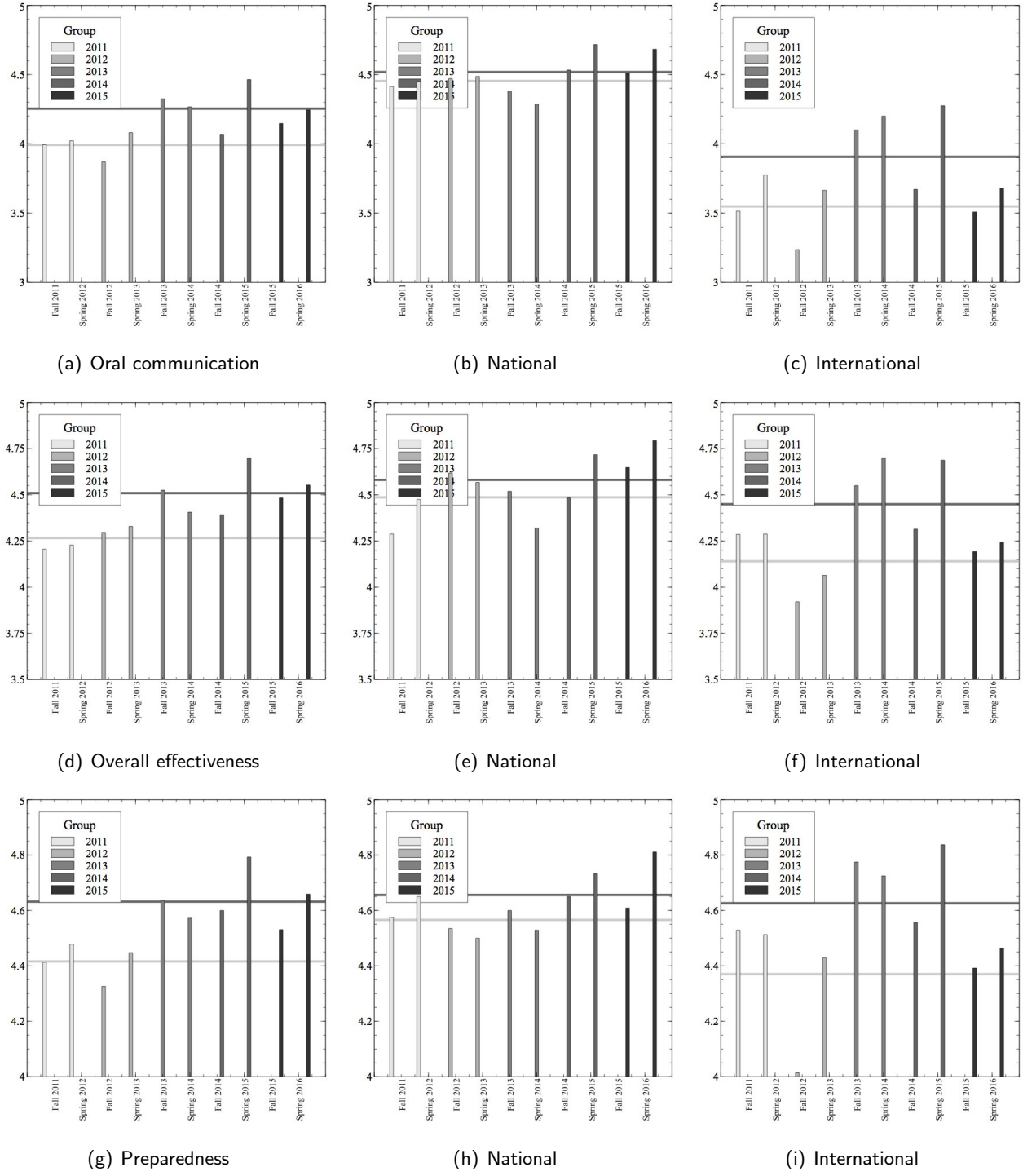


(d) Engaged students (e) National (f) International



(g) Explained concepts clearly (h) National (i) International

Figure 6: Each figure shows the means for every starting semester group, in their first fall and spring semester of teaching. The light horizontal line is the mean for the pre group(starting Fall 2011 and 2012) and the dark horizontal line is the mean for the post group (starting Fall 2013, 2014 and 2015).

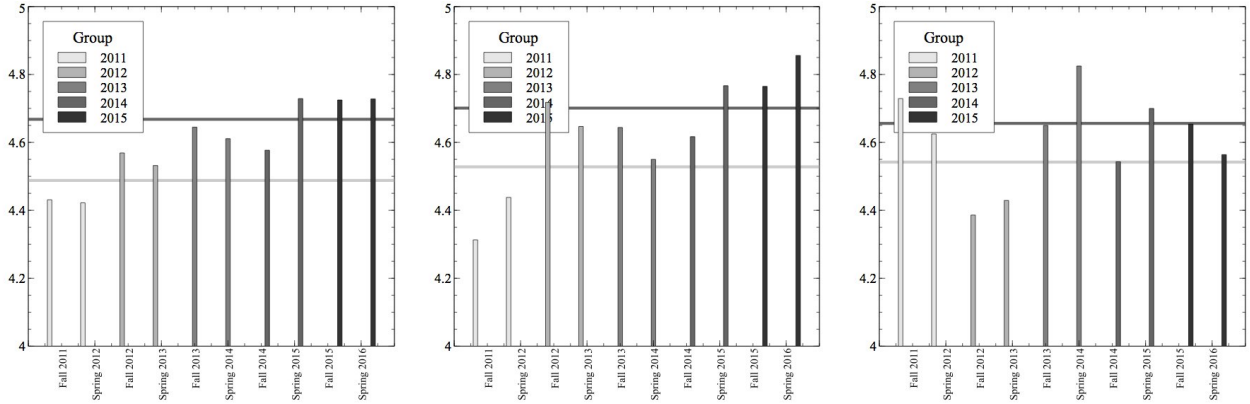


(a) Oral communication (b) National (c) International

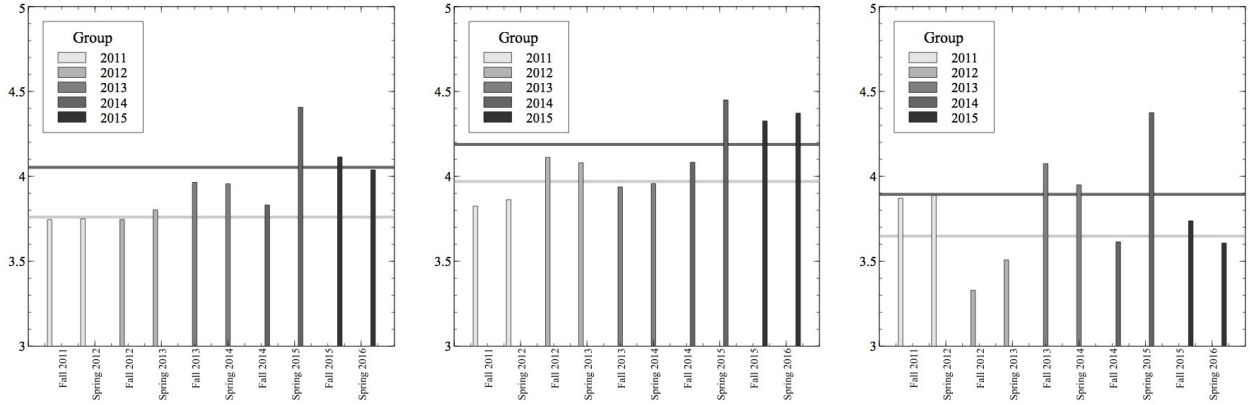
(d) Overall effectiveness (e) National (f) International

(g) Preparedness (h) National (i) International

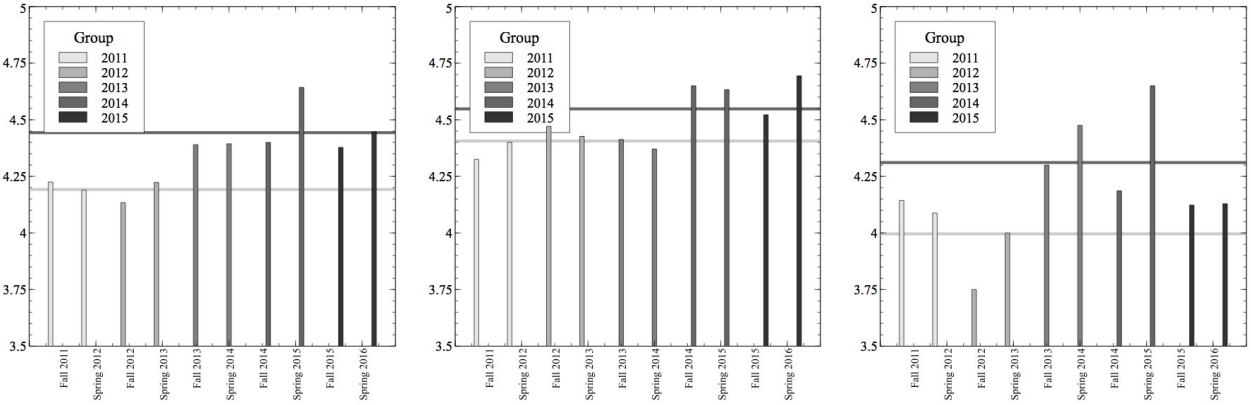
Figure 7: Each figure shows the means for every starting semester group, in their first fall and spring semester of teaching. The light horizontal line is the mean for the pre group(starting Fall 2011 and 2012) and the dark horizontal line is the mean for the post group (starting Fall 2013, 2014 and 2015).



(a) Respect for students (b) National (c) International



(d) Stimulated interest (e) National (f) International



(g) Written communication (h) National (i) International

Figure 8: Each figure shows the means for every starting semester group, in their first fall and spring semester of teaching. The light horizontal line is the mean for the pre group(starting Fall 2011 and 2012) and the dark horizontal line is the mean for the post group (starting Fall 2013, 2014 and 2015).

7. First year multiple independent variables tests

In the previous sections we have studied the effect of the different independent variables, specially the pre and post groups, on the scores of the TAOS survey for all the different questions of it. We have seen their individual influence on the scores and now we want to see a more overall result, by looking their effect at the same time. To do that we will perform a two-way ANOVA between the pre-post and national-international variables and between the pre-post and Traditional-M&I variables. After that, we will do a multilinear regression between all three independent variables, not looking for a perfect fit, but to see which coefficients are more important in each case. We will repeat that last analysis with the starting semester as independent variable instead the pre (starting 2011 and 2012) and post (starting 2013, 2014 and 2015) groups to see if there are some differences.

7.1 Two-Way ANOVA

As we have explained before, this test is the extension of the One-Way ANOVA for the case where there are two independent variables instead of only one. The assumptions are the same: independence of variables and independence of data inside the groups (in that case is fulfilled as we do different tests for the fall and spring semester), equality of variances and normality of the data. These two last conditions are not well fulfilled in our case, but as our objective in this section is to see the different effects of the different independent variables and the overall relations between them, this test is good enough. Moreover, in the analysis by semesters section, we have seen that although our data does not fulfill the required assumptions, the results where very similar to those obtained using the proper tests.

In the tables 15, 16, 17 and 18 we can see the effect of the different variables using this test, focusing on the pre-post variable. We can see how the national-international variable is almost always statistically significant (specially in the fall semester) and that the pre-post variable is statistically significant in the majority of the questions of the spring semester while in the fall semester is not that much important, as we have been seeing along all the analysis. The Traditional-M&I variable behaves differently, as it is clearly statistically significant in some questions while in others it is clearly not. This make sense as the two courses are different and some aspects require more skills and knowledge in one course than on the other. A quick conclusion that we can extract from this analysis is that the training program improves the performance of the GTAs in all the aspects that we study in a consistent way, meaning that the effect of the program is noticeable in almost all the categories and becomes more important in the spring semester of teaching. On the other hand, the nationality of the GTAs is probably the most determining factor on the GTAs performance, probably due to the language knowledge.

7.2 Multilinear regression

With this analysis we want to study the effect of the three independent variables at the same time. We could have performed a three-way ANOVA but we decided to do this as we are looking for a qualitatively overview of the effect of the training program, nationality and type of course on the TAOS scores. As we have said before, we are not looking for a perfect fit and we will only focus on the slope of the line (or regression coefficient) to see how important is the effect of the different variables. In principle the results obtained should agree with the ones obtained in the previous section, which showed a consistent influence of the training program (pre post groups or starting semester of teaching) a big influence of the nationality and some differences between the type of course in some specific questions. Before looking to the results

	Pre		Post				
Fall semester	**Nat**	**Int**	**Nat**	**Int**	Pre-Post	Nat-Int	Pre-Post*Nat-Int
Approachability	4.512	4.195	4.604	4.475	0.036	0.012	0.287
Attitude about teaching	4.460	4.224	4.520	4.388	0.151	0.019	0.504
Classroom management	4.484	4.219	4.576	4.325	0.227	0.002	0.930
Concept familiarity	4.600	4.386	4.780	4.596	0.003	0.002	0.815
Engaged students	4.404	4.043	4.536	4.200	0.084	0.001	0.877
Explained concepts clearly	4.408	3.657	4.489	3.988	0.030	0.001	0.184
Oral communication	4.452	3.329	4.467	3.654	0.075	0.001	0.104
Overall effectiveness	4.512	4.043	4.580	4.287	0.065	0.001	0.294
Preparedness	4.548	4.186	4.611	4.504	0.012	0.002	0.091
Respect for students	4.588	4.500	4.702	4.621	0.099	0.233	0.963
Stimulated interest	4.020	3.510	4.156	3.758	0.060	0.001	0.577
Written communication	4.424	3.881	4.500	4.171	0.028	0.001	0.196

Table 15: Means of the pre and post groups of the fall semester, divided between national and international; p-values of the two-way ANOVA for the pre-post, national-international and for the relation between independent variables.

	Pre		Post				
Spring semester	**Nat**	**Int**	**Nat**	**Int**	Pre-Post	Nat-Int	Pre-Post*Nat-Int
Approachability	4.596	4.332	4.642	4.481	0.223	0.009	0.552
Attitude about teaching	4.357	4.186	4.489	4.458	0.029	0.273	0.451
Classroom management	4.470	4.386	4.655	4.512	0.041	0.134	0.688
Concept familiarity	4.674	4.495	4.789	4.669	0.009	0.007	0.593
Engaged students	4.435	4.177	4.568	4.312	0.161	0.008	0.997
Explained concepts clearly	4.383	3.968	4.592	4.231	0.009	0.001	0.767
Oral communication	4.474	3.705	4.542	3.942	0.121	0.001	0.388
Overall effectiveness	4.535	4.145	4.608	4.450	0.060	0.007	0.247
Preparedness	4.552	4.459	4.695	4.619	0.016	0.176	0.887
Respect for students	4.574	4.500	4.729	4.646	0.020	0.223	0.945
Stimulated interest	4.004	3.652	4.232	3.896	0.020	0.001	0.934
Written communication	4.417	4.033	4.566	4.342	0.002	0.001	0.278

Table 16: Means of the pre and post groups of the spring semester, divided between national and international; p-values of the two-way ANOVA for the pre-post, national-international and for the relation between independent variables.

	Pre		Post				
Fall semester	**Trad**	**M&I**	**Trad**	**M&I**	Pre-Post	Trad-M&I	Pre-Post*Trad-M&I
Approachability	4.507	4.328	4.662	4.485	0.078	0.045	0.991
Attitude about teaching	4.293	4.390	4.531	4.432	0.085	0.989	0.231
Classroom management	4.533	4.300	4.590	4.415	0.306	0.016	0.725
Concept familiarity	4.720	4.421	4.734	4.703	0.020	0.010	0.036
Engaged students	4.227	4.238	4.486	4.370	0.034	0.565	0.485
Explained concepts clearly	3.920	4.159	4.386	4.263	0.011	0.605	0.105
Oral communication	3.753	4.048	4.276	4.118	0.028	0.608	0.091
Overall effectiveness	4.333	4.283	4.534	4.437	0.059	0.431	0.804
Preparedness	4.593	4.317	4.628	4.535	0.088	0.013	0.213
Respect for students	4.593	4.510	4.769	4.605	0.064	0.090	0.576
Stimulated interest	3.740	3.817	4.059	3.987	0.032	0.978	0.511
Written communication	4.200	4.214	4.428	4.355	0.042	0.744	0.631

Table 17: Means of the pre and post groups of the fall semester, divided between traditional and M&I; p-values of the two-way ANOVA for the pre-post, traditional-M&I and for the relation between independent variables.

	Pre		Post				
Spring semester	**Trad**	**M&I**	**Trad**	**M&I**	Pre-Post	Trad-M&I	Pre-Post*Trad-M&I
Approachability	4.494	4.456	4.696	4.495	0.143	0.144	0.320
Attitude about teaching	4.261	4.288	4.546	4.429	0.023	0.624	0.435
Classroom management	4.550	4.344	4.704	4.524	0.027	0.011	0.862
Concept familiarity	4.611	4.576	4.808	4.695	0.005	0.185	0.484
Engaged students	4.311	4.332	4.538	4.413	0.113	0.589	0.450
Explained concepts clearly	4.106	4.224	4.488	4.416	0.004	0.815	0.330
Oral communication	4.000	4.172	4.369	4.250	0.062	0.824	0.221
Overall effectiveness	4.300	4.388	4.615	4.495	0.044	0.875	0.315
Preparedness	4.578	4.464	4.704	4.637	0.018	0.149	0.707
Respect for students	4.578	4.500	4.785	4.634	0.009	0.080	0.574
Stimulated interest	3.772	3.875	4.146	4.061	0.009	0.935	0.372
Written communication	4.239	4.229	4.512	4.450	0.003	0.656	0.746

Table 18: Means of the pre and post groups of the spring semester, divided between traditional and M&I; p-values of the two-way ANOVA for the pre-post, traditional-M&I and for the relation between independent variables.

we have to consider that this analysis has been performed assigning the following values to the variables: pre = 1, post = 2, national = 1, international = 2, Traditional = 1, M&I = 2; so the coefficients in the first analysis will be equally comparable as the differences in the independent variable values are the same. The results of this analysis is shown on the tables 19 and 20. On the second analysis we have the following values for the starting semester of teaching: 11, 12, 13, 14 and 15, so in that case the regression coefficient of this variable will be approximately 4 times smaller compared to the other ones, and it has to be corrected in order to compare the three of them. We have also performed the analysis changing the starting semester values and the results can be seen on tables 21 and 22. We can see how the results are almost the same if we consider the pre-post variable or the starting semester of teaching. We can also note that the effect of the training program is more or less the same among the different questions, being always positive towards the post group or the lasts starting semesters. The nationality is in some cases very important, especially to those related to the language: explain concepts clearly, oral and written communication; and apart of these it is also a very influent factor specially in the fall semester. We also see the same behavior that we have seen before with the type of course, being usually the Traditional course the one with better scores and having some categories with a clear difference (such as approachability, classroom management, preparedness or respect for students) and some other categories without a noticeable difference (such as explain concepts clearly, oral communication or stimulated interest). Moreover, in this analysis we also see that the effect of the training program increases on the spring semester, as we have been noticing on the previous results.

In conclusion we can say that this analysis shows us how the CETL 8000 PH has a persistent effect on all the categories and that effect is more important on the second semester of teaching, having then a good evolution on the GTAs.

Fall semester	**Pre-Post**	**National-International**	**Traditional-M&I**
Approachability	0.136	-0.185	-0.185
Attitude about teaching	0.097	-0.164	-0.033
Classroom management	0.067	-0.242	-0.206
Concept familiarity	0.163	-0.177	-0.138
Engaged students	0.142	-0.336	-0.083
Explained concepts clearly	0.174	-0.584	-0.013
Oral communication	0.135	-0.926	-0.028
Overall effectiveness	0.133	-0.357	-0.094
Preparedness	0.129	-0.188	-0.168
Respect for students	0.115	-0.090	-0.138
Stimulated interest	0.178	-0.428	-0.034
Written communication	0.130	-0.391	-0.057

Table 19: Coefficients of the multilinear regression with pre-post, national-international and traditional M&I as independent variables, for the fall semester.

Spring semester	Pre-Post	National-International	Traditional-M&I
Approachability	0.091	-0.177	-0.126
Attitude about teaching	0.196	-0.058	-0.056
Classroom management	0.162	-0.088	-0.186
Concept familiarity	0.141	-0.124	-0.075
Engaged students	0.123	-0.221	-0.055
Explained concepts clearly	0.241	-0.366	0.024
Oral communication	0.145	-0.649	0.032
Overall effectiveness	0.174	-0.228	-0.024
Preparedness	0.148	-0.061	-0.083
Respect for students	0.159	-0.061	-0.118
Stimulated interest	0.242	-0.316	0.004
Written communication	0.223	-0.270	-0.029

Table 20: Coefficients of the multilinear regression with pre-post, national-international and traditional M&I as independent variables, for the spring semester.

Fall semester	Starting Semester	SS corrected	Nat-Int	Traditional-M&I
Approachability	0.053	0.213	-0.191	-0.189
Attitude about teaching	0.033	0.130	-0.169	-0.036
Classroom management	0.032	0.128	-0.245	-0.207
Concept familiarity	0.033	0.133	-0.189	-0.147
Engaged students	0.053	0.210	-0.343	-0.087
Explained concepts clearly	0.046	0.183	-0.596	-0.021
Oral communication	0.031	0.125	-0.936	-0.035
Overall effectiveness	0.041	0.165	-0.365	-0.099
Preparedness	0.020	0.080	-0.199	-0.176
Respect for students	0.047	0.189	-0.095	-0.140
Stimulated interest	0.072	0.290	-0.436	-0.038
Written communication	0.039	0.158	-0.399	-0.062

Table 21: Coefficients of the multilinear regression with starting semester, starting semester corrected, national-international and traditional M&I as independent variables, for the fall semester.

Spring semester	Starting Semester	SS corrected	Nat-Int	Traditional-M&I
Approachability	0.032	0.126	-0.183	-0.128
Attitude about teaching	0.066	0.262	-0.070	-0.060
Classroom management	0.060	0.239	-0.098	-0.190
Concept familiarity	0.035	0.141	-0.134	-0.077
Engaged students	0.044	0.176	-0.228	-0.058
Explained concepts clearly	0.065	0.262	-0.381	0.020
Oral communication	0.047	0.189	-0.658	0.029
Overall effectiveness	0.065	0.260	-0.239	-0.029
Preparedness	0.047	0.186	-0.070	-0.085
Respect for students	0.060	0.238	-0.071	-0.122
Stimulated interest	0.075	0.301	-0.331	-0.001
Written communication	0.071	0.284	-0.283	-0.033

Table 22: Coefficients of the multilinear regression with starting semester, starting semester corrected, national-international and traditional M&I as independent variables, for the spring semester.

8. All data results student by student

In this final analysis we want to study the evolution of the GTAs through all their semesters of teaching. The first obvious problem is that if we repeat the analysis done before we would violate the independence assumptions and thus the results wouldn't be reliable. Another problem that we have is that a lot of the GTAs only teach their first year as Graduate Students and this makes very difficult to see their evolution as teachers. So due to the lack of data for the non-first year GTAs and the impossibility to perform certain analysis with this data, we decided to study the distributions of the difference (subtraction) between teaching semesters of the GTAs. That means that for every GTA we computed the difference between one semester of teaching and the previous one (of the same GTA) in order to see if the GTAs had a positive or negative progression. Our main objective with this test is to study their evolution considering if they had done the training program CETL 8000 PH (Post group: starting 2013, 2014 and 2015) or not (Pre group: starting 2011 and 2012).

This analysis won't show us which group has better scores on the TAOS survey (that has already been discussed in the previous sections), but it will show us how the GTAs of each group evolve through their semesters of teaching.

The tables 23 and 24 show the means, standard deviations and kurtosis coefficients (the kurtosis coefficient indicates the concentration of data around the mean, so a higher kurtosis coefficient will indicate a higher concentration around the mean which also indicates a more peaked distribution) of the distributions explained. The first table is done with only the data of the first year, that is to say that the distribution studied is only the difference between the first spring and the first fall semesters of teaching of the GTAs (i.e. we only look to the first year evolution). While the second table is done with all the data available (including the data of the first year), so it shows the results of the distribution of the difference of all the semesters in which the GTAs were teaching. The results of theses two tables will be similar as the majority of the data corresponds to the first year. We have reported the means to see the evolution itself of the GTAs' teaching, as if the mean is positive will indicate a general positive evolution (i.e. the GTAs have better scores every semester) and if the mean is negative it will indicate the opposite (the GTAs have worst scores every semester). The standard deviation and the kurtosis coefficient are calculated to study

the consistency of teaching, that is, if we have lower standard deviation and higher kurtosis coefficients it would mean that the variations of the scores are small and thus that the GTAs always teach the same way; while if we have the opposite (higher standard deviations and lower kurtosis coefficients) it could mean that the GTAs don't have a well defined way to teach and that could lead to higher variations on their scores.

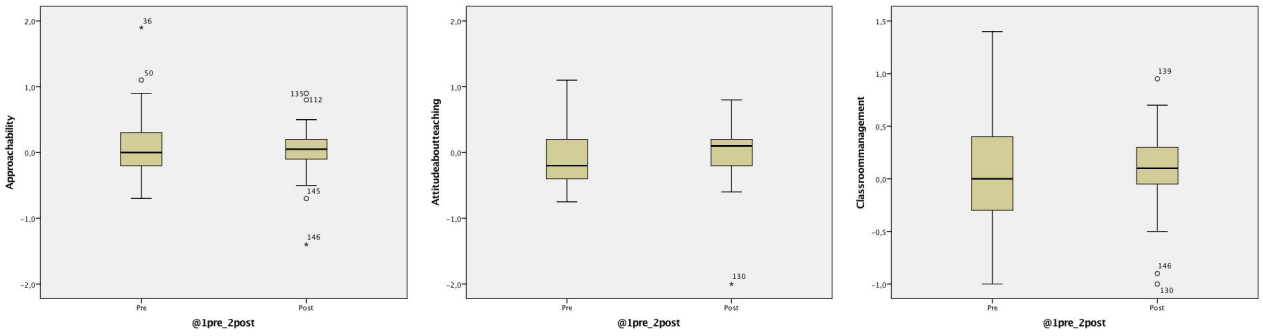
In the tables 23 and 24 we can see how the means of the post group are positive for all the categories while there are some from the pre group which are negatives, which indicates that the general evolution of the post group is always positive while the general evolution of the pre group is sometimes negative. The standard deviations of the post group of the first year evolution are all smaller than the ones of the pre group (except one case where they are almost equal), while in the all semesters evolution we have some cases where the standard deviation of the pre group is smaller. Finally, the kurtosis coefficients of the post group are usually higher that the ones of the pre group except for the last two categories which the pre group has clearly higher kurtosis coefficients.

We can also see in the figures 9 and 10 the boxplots of the distributions for the pre and post groups, and we can notice how the post group has more compact distributions that the pre group. We can also see how the median is equal or higher for the post group, which also indicates a better evolution for the GTAs of this group.

With all these results, we can conclude that in general the evolution of the GTAs of the post group (those who did the CETL 8000 PH) is better that the evolution of the GTAs of the pre group; as it is positive in all the categories and has in general less variations, which could indicate a more consistent way of teaching produced by a better knowledge of teaching techniques and, in general, by a better preparation.

FIRST YEAR	**Mean Pre**	**Mean Post**	**SD Pre**	**SD Post**	**Kurt. Pre**	**Kurt. Post**
Approachability	0.087	0.022	0.4969	0.3659	2.810	3.921
Attitudeaboutteaching	-0.084	0.014	0.3947	0.3981	0.157	12.114
Classroommanagement	0.063	0.114	0.4896	0.3763	0.669	1.223
Conceptfamiliarity	0.097	0.016	0.4667	0.3109	1.225	6.021
Engagedstudents	0.023	0.065	0.5325	0.4315	-0.457	3.593
Explainedconceptsclearly	0.103	0.164	0.5093	0.3645	-0.282	2.120
Oralcommunication	0.146	0.195	0.5585	0.3902	0.280	0.129
Overalleffectiveness	0.030	0.060	0.5011	0.4727	-0.101	12.068
Preparedness	0.116	0.090	0.4885	0.3340	1.245	1.304
Respectforstudents	-0.014	0.003	0.3573	0.2369	2.990	0.862
Stimulatedinterest	-0.033	0.116	0.7800	0.4463	14.462	1.011
Writtencommunication	-0.072	0.129	0.7204	0.3999	15.527	0.850

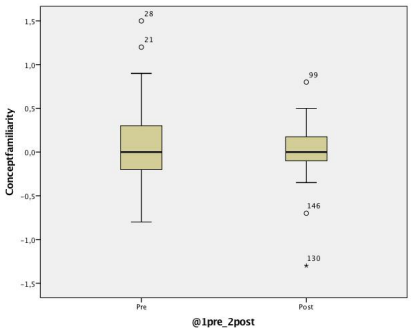
Table 23: Study of the evolution of the GTAs among the first year of teaching. Means, Standard Deviations and Kurtosis coefficients of the distributions of the differences between the first spring and first fall semesters, for the GTAs who taught in both semesters; divided by Pre and Post groups.



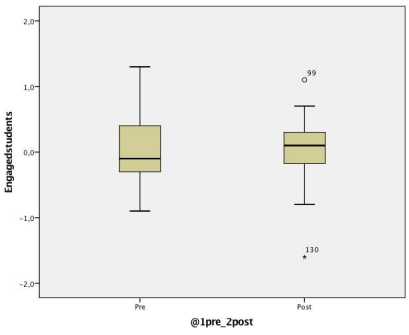
(a) Approachability

(b) Attitude about teaching

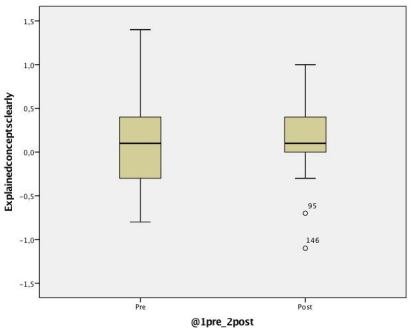
(c) Classroom management



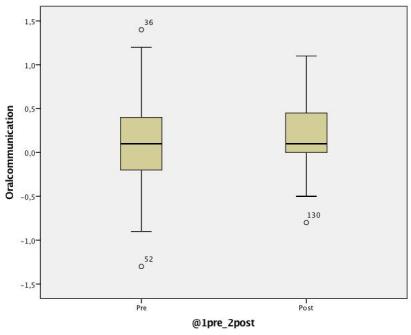
(d) Concept familiarity



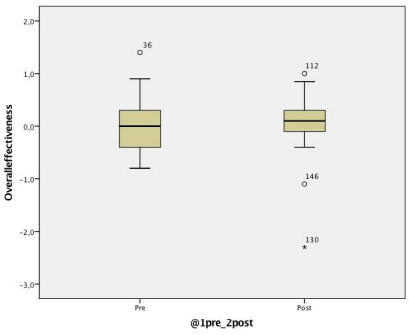
(e) Engaged students



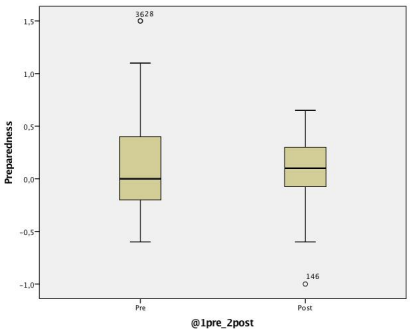
(f) Explained concepts clearly



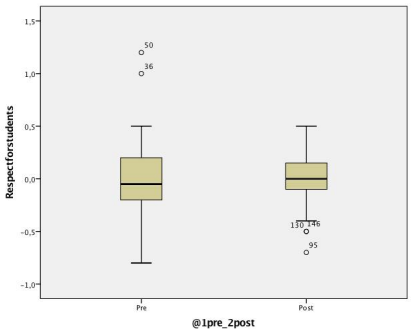
(g) Oral communication



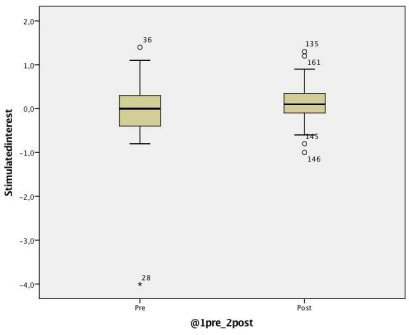
(h) Overall effectiveness



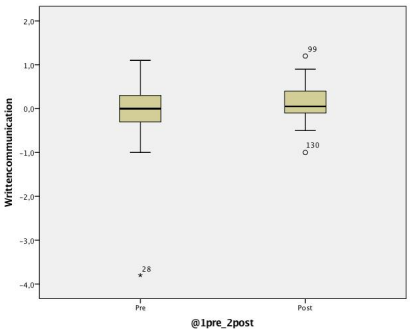
(i) Preparedness



(j) Respect for students

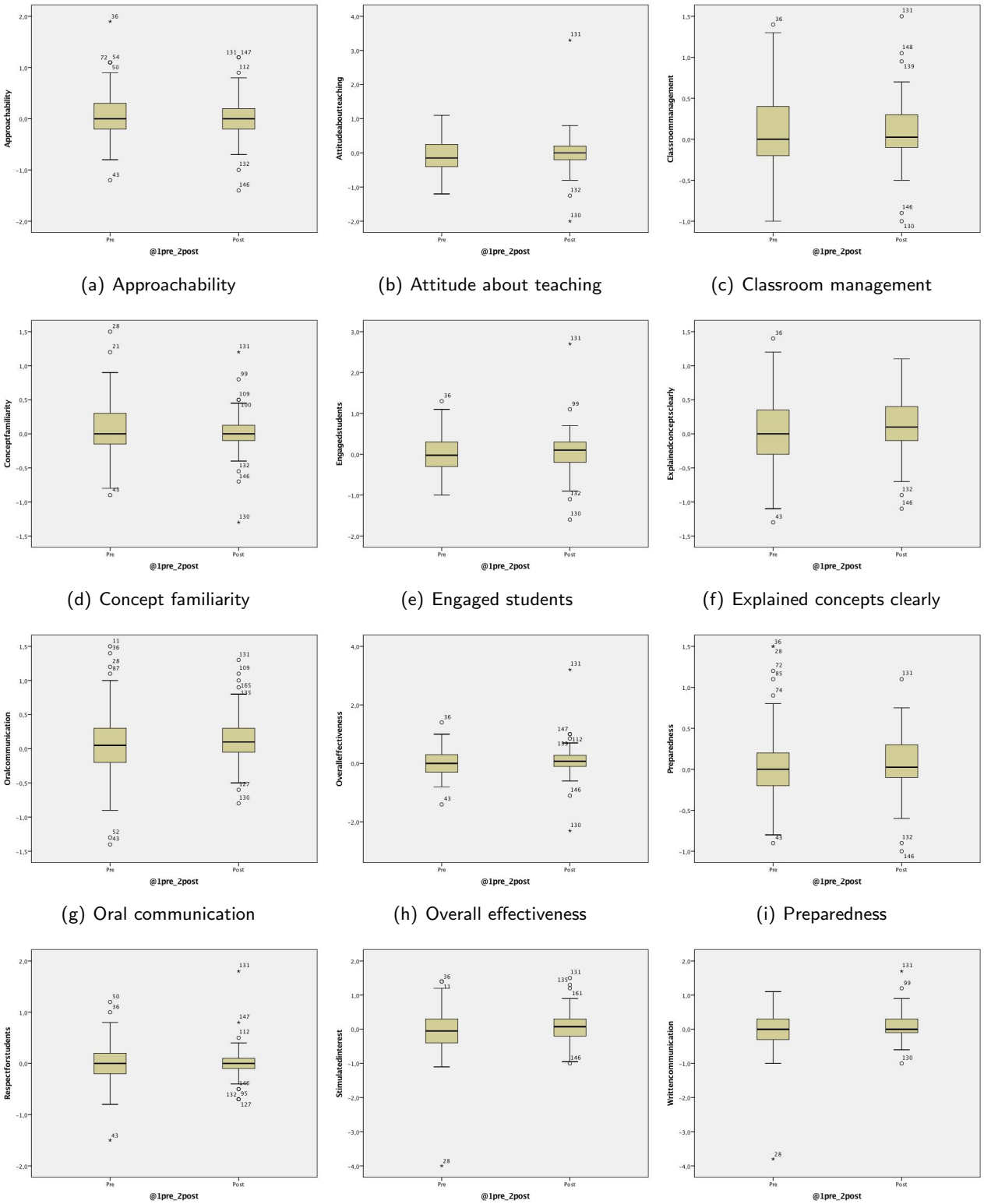


(k) Stimulated interest



(l) Written communication

Figure 9: Boxplots of the distributions of the differences between the first spring and fall semesters of teaching, divided by pre and post groups.



(a) Approachability (b) Attitude about teaching (c) Classroom management

(d) Concept familiarity (e) Engaged students (f) Explained concepts clearly

(g) Oral communication (h) Overall effectiveness (i) Preparedness

(j) Respect for students (k) Stimulated interest (l) Written communication

Figure 10: Boxplots of the distributions of the differences between consecutive semesters of teaching, divided by pre and post groups.

ALL SEMESTERS	Mean Pre	Mean Post	SD Pre	SD Post	Kurt. Pre	Kurt. Post
Approachability	0.059	0.016	0.4835	0.4025	1.729	3.003
Attitudeaboutteaching	-0.071	0.018	0.4460	0.5600	0.051	17.011
Classroommanagement	0.056	0.089	0.4346	0.4004	0.609	1.964
Conceptfamiliarity	0.077	0.016	0.4087	0.3174	1.408	5.519
Engagedstudents	0.012	0.061	0.5062	0.5374	-0.319	7.733
Explainedconceptsclearly	0.043	0.118	0.5383	0.3915	-0.040	1.266
Oralcommunication	0.064	0.155	0.5224	0.3864	0.865	0.743
Overalleffectiveness	0.023	0.076	0.4987	0.5639	0.056	15.708
Preparedness	0.072	0.068	0.4472	0.3515	1.426	1.439
Respectforstudents	-0.009	0.015	0.3825	0.3255	2.753	11.888
Stimulatedinterest	-0.024	0.090	0.6897	0.4812	10.887	0.861
Writtencommunication	-0.001	0.104	0.5851	0.4156	18.163	2.638

Table 24: Study of the evolution of the GTAs among their years of teaching. Means, Standard Deviations and Kurtosis coefficients of the distributions of the differences between consecutive semesters; divided by Pre and Post groups.

Part IV

Discussions and conclusion

In the results section we have seen a lot of analysis and we have extracted some results from them, but they have been general and without focusing on each question of the TAOS survey. Now we want to analyze in detail the results for each category and try to extract valuable information in order to be able to modify or adapt the future training program in the cases where it could be possible.

Before starting the analysis with detail, we should remember some overall conclusions that we have been seeing during the tests. The scores of the national and post groups have been always higher that the scores from the international and pre groups, and when there are noticeable differences between the Traditional and the M&I courses, the score of the Traditional course is higher. We have also seen that, in general, the differences in means (in the first year of teaching of the GTAs) of the pre and post groups are higher (and then more statistically significant different) in the spring semester than in the fall semester.

- **Approachability:** The GTA was accessible for assistance during the course.

  Even though it seems something that shouldn't depend on the nationality of the GTA, there are significant differences between national and international GTAs, and also between the Traditional and M&I courses. The CETL 8000 PH has the same effect in mean increase in all the subgroups (national-international, Traditional-M&I) of the GTAs. This increase could be due to the importance given by the training program to pay attention to the students and help them both in class and outside class, and it seems to affect to all the GTAs equally.

- **Attitude about teaching:** The GTA's attitude about their teaching role in this course.

  Even though there are statistically significant differences between the national and international GTAs, is one of the categories where the difference is smaller. Between the Traditional and M&I courses there are no significant differences although, as usual, the Traditional scores are higher. The difference between pre and post is not statistically significant in the fall semester but in the spring

semester it is clearly significant, and this behavior is repeated in all the subgroups (except for the Traditional GTAs where the difference is always statistically significant). This is one of the categories where the effect of the training program becomes much more important in the second semester of teaching, and it would be interesting to have more data to analyze properly the evolution of the GTAs through more semesters.

- **Classroom management:** The GTA's management of classroom/lab environment.

  In this category, we have important differences between national and international GTAs as well as between Traditional and M&I courses. A noticeable fact about the classroom management is that the scores are higher in the spring semester without regarding if the GTAs had done the training program. This could be due to the self-learning of the GTAs, which is also important in other categories as the concept familiarity. The effect of the training program is, as usual, more observable in the spring semester, where the GTAs who had done the CETL 8000 PH have a bigger increase in their scores. This makes us think that, although there is a self-learning improvement from the GTAs, this improvement is considerably bigger for those GTAs who have had a proper training program.

- **Concept familiarity:** The GTA's familiarity with course concepts.

  As we just said, this is also a category where the GTAs improve by themselves, as the contents are more or less the same from one semester to the other and they should have a better knowledge every time they teach the class. This is reflected in the evolution means of the pre and post groups in the tables first year and ALL year, where we can see how the pre group means are higher than the ones from the post group, due to the fact that they have more room for improvement and that they know better the contents every time they teach. In this category, there are big differences between national and international GTAs, but the improvement of the international GTAs is also bigger in the spring semester. The effect of the training program is very important in both semesters, and this fact is very significant as the concept familiarity is one of the main bases of the CETL 8000 PH.

- **Engaged students:** The GTA actively engaged students, for example with questions, participation, group work, etc.

  As with the majority of categories, there are significant differences between the national and international GTAs, but this differences are much smaller when comparing the Traditional and M&I courses. The training program has a similar effect on all the subgroups and the improvement on the scores is more or less the same in both semesters. The main peculiarity of this category is that the impact of the CTEL 8000 PH is bigger in the Traditional course than in the M&I, probably due to the fact that the tips given by the training program can be more easily applied in a recitation class rather than in a laboratory.

- **Explained concepts clearly:** The GTA explained course concepts clearly.

  This is one of the categories where the difference between national and international GTAs is bigger, probably due to the better knowledge of the language. On the other hand, there are no differences between the Traditional and M&I courses. The international GTAs have an important improvement between the first and second semester without regard of the training program, although the scores are much higher for the ones of the post group. The effect of the training program is very velar on the spring semester of all the subgroups, while in the fall semester it is only statistically significant for the GTAs of the Traditional course.

- **Oral communication:** The GTA's oral communication skills.

This is the category where the difference between the national and international GTAs is bigger, obviously due to the fact that the national GTAs have a better knowledge of the language. Another think to take into account is that the GTAs from the School of Physics are the ones who are required better scores in the English exams, and that the language requirements, for international GTAs, have not changed since before 2010, so the English level of the international GTAs should be the same for those of the pre and post groups. On the other hand, there are no significant differences between the Traditional and M&I courses, although the Traditional course has recitation lessons which could affect negatively to the scores of those GTAs with more difficulties with the language. The training program has a significant effect on the GTAs, but this is more important on the international students as they have more room for improvement. The other important fact is that the improvement on the M&I course is very small compared to the improvement of the Traditional course (where it is clearly significant). That could be due to the recitation lessons that the Traditional course has and the M&I don't.

- **Overall effectiveness:** Considering everything, the GTA was an effective GTA.

  This category should reflect a summary of all the performance of the GTAs. In that case we can say that the national GTAs have much higher scores than the international ones, and that there are no important differences between the two type of courses. If we look to the pre and post groups, we can see how the differences are more significant in the spring semester (as it happens with the majority of categories) and that the training program affects with a similar impact to all the subgroups.

- **Preparedness:** The GTA's level of preparation.

  As usual, the national GTAs have better scores than the international, even though there are no previous reasons to think that a national GTA should have a better preparation. The difference between Traditional and M&I is more important on the fall semester, what makes us think that the GTAs can improve in this aspect by self-learning. This fact can also be observed looking to the scores of the pre group between the fall and spring semester, where we can see that there is an important improvement. This improvement between the fall and spring semester is also important on the post group, and the results show how the differences become bigger on the spring semester, despite the higher scores of the pre group. So we can say that even the preparation is something that it's improved by the experience, a proper training not only has repercussions in the first semester of teaching, but also helps to improve more this aspect every semester.

- **Respect for students:** The GTA's respect for their students.

  This is the only category where the differences between national and international GTAs are not statistically significant in the fall semester. This makes sense as the respect for the students shouldn't depend on the nationality or the course type, but as we can see in the results, there are important differences on the spring semester and also between course type on the fall semester. The training program affects to this aspect the same way as to many of the categories analyzed, as there are no statistically significant differences between the pre and post groups on the fall semester but they are clearly significant on the spring semester. One important fact of this category is that the pre group does not improve their scores from one semester to the other, meaning that, without the proper training, the GTAs do not improve their performance with time.

- **Stimulated interest:** The GTA stimulated my interest in the subject matter.

  Like other aspects like the oral communication, there is a big difference between the national and the international GTAs, while there is not a clear difference between the Traditional and the M&I

courses. A peculiarity of this category is that the data is more or less normally distributed (specially for the pre group), and it is the only one which is like this. The effect of the training program is also clear in all the subgroups in a similar way, specially in the spring semester. So we cold conclude that the training program has had a consistent effect on all the GTAs in this aspect of the teaching performance.

- **Written communication:** The GTA's written communication skills.

  As with the previous category, the differences between national and international GTAs are clear, while there are no differences between the traditional and M&I courses. The training program has also the same effect than with the stimulated interest, some differences in the fall semester (but not statistically significant in all the subgroups) and clearly significant in the spring semester. In that case there is a subgroup where the written communication clearly is improved from the fall to the spring semester without the need of a training program, it is the case of the international GTAs who, like in previous cases, have more room for improvement and thus, their scores are higher in the spring semester.

Conclusion

The main objective of this thesis was to decide if the training program CETL 8000 PH had a significant effect on the students. To do that, we have analyzed the scores of the TAOS surveys in many different ways, trying to extract as much information as we could. With all the results obtained and the discussions in the previous section, we can conclude that indeed, the CETL 8000 PH has had a positive statistically significant effect on the students of the School of Physics of the Georgia Tech Institute of Technology. That effect is more or less important in the different categories (or questions of the survey) but in all the cases, and dividing the GTAs by different ways (national-international, Traditional-M&I), the effect is always positive, meaning that the scores of the GTAs who had done the training program are, in mean, always higher than the ones from the GTAs who didn't do the training program. Moreover, we have seen how in many cases those improvements were statistically significant. This improvement in the teaching performance of the GTAs indicates that the CETL 8000 PH is noticeably useful for the students, as well as for the GTAs. All the information obtained from these analyses will be used to improve the training program itself, trying to give to the new GTAs the proper preparation in order to become better teaching assistants and so allowing the students to receive a better education.

References

[1] Michael J Campbell. Teaching non-parametric statistics to students in health sciences. In *ICOTS 7*. IASE, ISI Salvador, Bahia, Brazil, 2006.

[2] Ruth W Chabay. *Matter and Interactions: Modern Mechanics/Electric and Magnetic Interactions*. John Wiley & Sons, 2015.

[3] Kenneth A Feldman. Grades and college students' evaluations of their courses and teachers. *Research in Higher Education*, 4(1):69–111, 1976.

[4] David F Feldon, James Peugh, Briana E Timmerman, Michelle A Maher, Melissa Hurst, Denise Strickland, Joanna A Gilmore, and Cindy Stiegelmeyer. Graduate students teaching experiences improve their methodological research skills. *Science*, 333(6045):1037–1039, 2011.

[5] Learner-Centered Task Force. Learner-centered teaching and education at usc: A resource for faculty.

[6] John A Gilreath and Timothy F Slater. Training graduate teaching assistants to be better undergraduate physics educators. *Physics Education*, 29(4):200, 1994.

[7] Randall Dewey Knight. Physics for scientists and engineers: a strategic approach 3 rd ed. 2013.

[8] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

[9] Frances Lawrenz et al. Training the teaching assistant. *Journal of College Science Teaching*, 22(2):106–9, 1992.

[10] Howard Levene. Robust tests for equality of variances1. *Contributions to probability and statistics: Essays in honor of Harold Hotelling*, 2:278–292, 1960.

[11] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[12] LD Muhlestein and B DeFacio. Teaching graduate teaching assistants to teach. *American Journal of Physics*, 42(5):384–386, 1974.

[13] Nornadiah Mohd Razali, Yap Bee Wah, et al. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, 2(1):21–33, 2011.

[14] SS Shaphiro and MB Wilk. An analysis of variance test for normality. *Biometrika*, 52(3):591–611, 1965.

[15] FB Stumpf. A course for graduate preparation for teaching. *American Journal of Physics*, 39(10):1223–1225, 1971.

[16] Bernard L Welch. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947.

[17] BL Welch. On the comparison of several mean values: an alternative approach. *Biometrika*, 38(3/4):330–336, 1951.

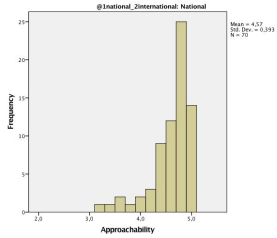
[18] Thomas K Worthen. The frustrated gta: A qualitative investigation identifying the needs within the graduate teaching assistant experience. 1992.

A. Appendix 1

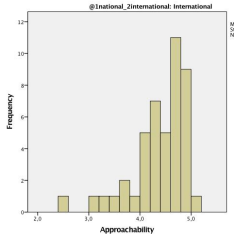
In this appendix we will see if the distributions of the different groups have similar shapes in order to be able to perform the Mann-Whitney test with all the assumptions fulfilled. To do that we will compare the histograms of the different distributions and see that they have, more or less, the same shape.

A.1 Comparing different independent variables assumption

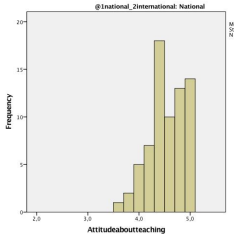
In this section we are looking to the shapes of the distributions divided by National-International and Traditional-M&I, without taking into account the pre and post groups. Figures 11 and 12 show the histograms of the fall semester of the National and International GTAs and figures 13 and 14 show the same histograms for the spring semester. Figures 15, 16, 17 and 18 show the histograms of the fall and spring semesters for the Traditional and M&I courses.



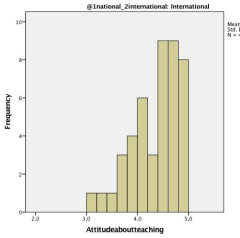
(a) Approachability National



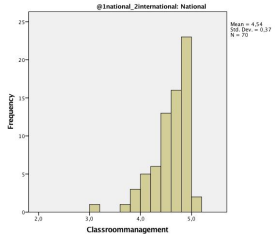
(b) Approachability International



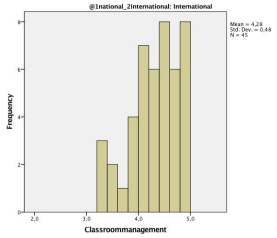
(c) Attitude about teaching National



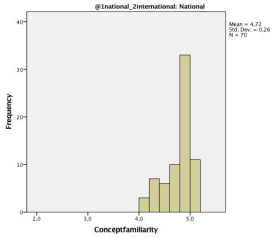
(d) Attitude about teaching International



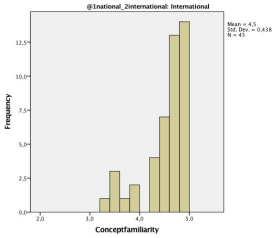
(e) Classroom management National



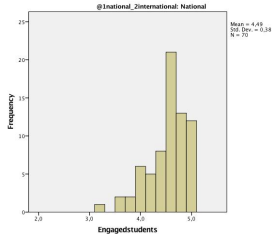
(f) Classroom management International



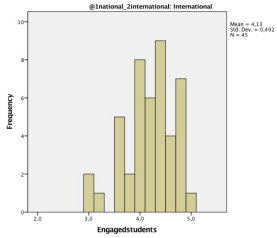
(g) Concept familiarity National



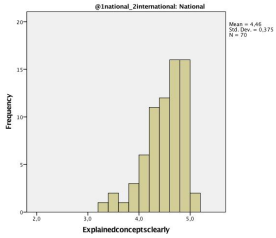
(h) Concept familiarity International



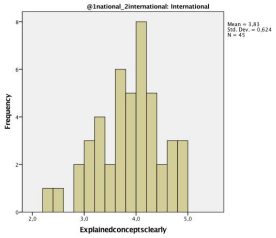
(i) Engaged students National



(j) Engaged students International

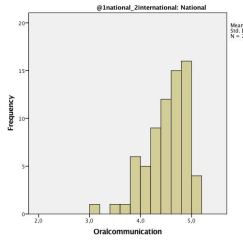


(k) Explained concepts clearly National

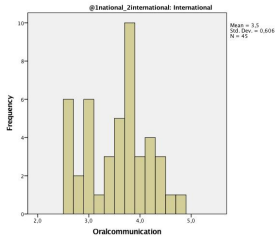


(l) Explained concepts clearly International

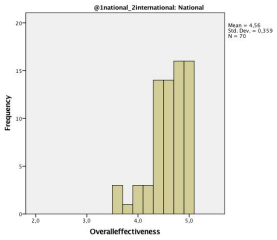
Figure 11: Histograms of the distributions of the national and international GTAs for the fall semester of teaching.



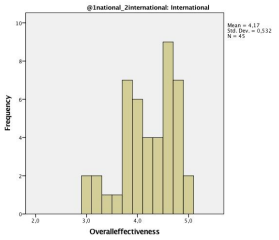
(a) Oral communication National



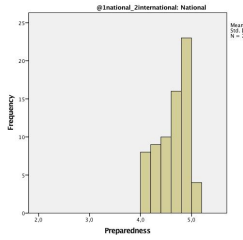
(b) Oral communication International



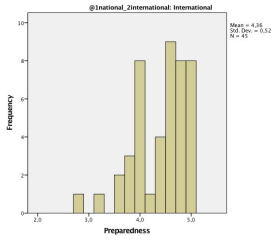
(c) Overall effectiveness National



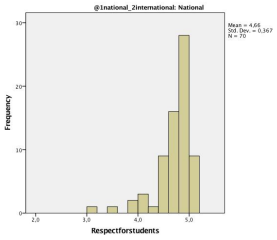
(d) Overall effectiveness International



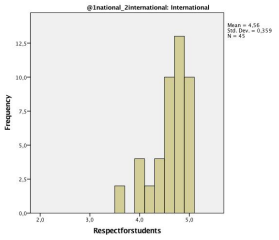
(e) Preparedness National



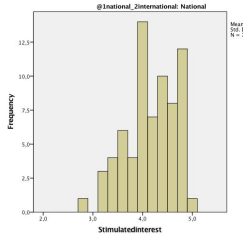
(f) Preparedness International



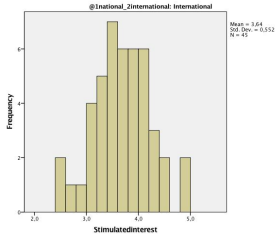
(g) Respect for students National



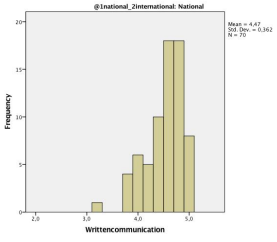
(h) Respect for students International



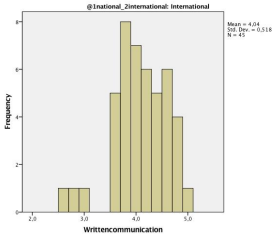
(i) Stimulated interest National



(j) Stimulated interest International

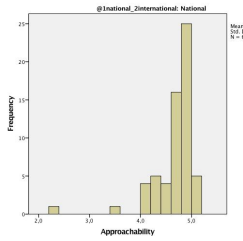


(k) Written communication National

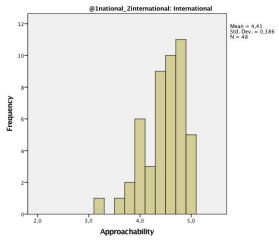


(l) Written communication International

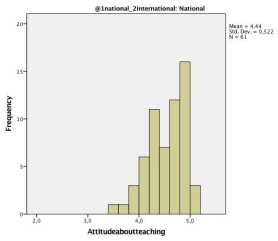
Figure 12: Histograms of the distributions of the national and international GTAs for the fall semester of teaching.



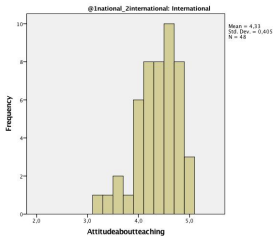
(a) Approachability National



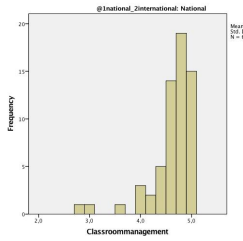
(b) Approachability International



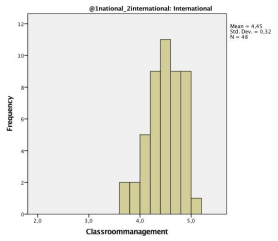
(c) Attitude about teaching National



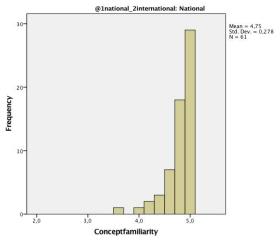
(d) Attitude about teaching International



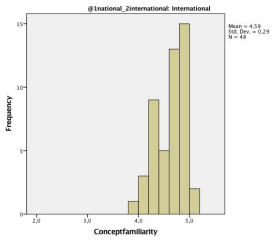
(e) Classroom management National



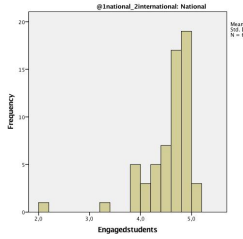
(f) Classroom management International



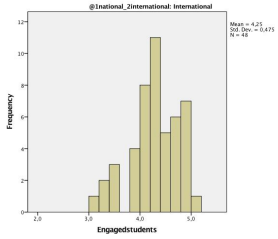
(g) Concept familiarity National



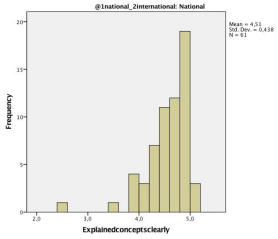
(h) Concept familiarity International



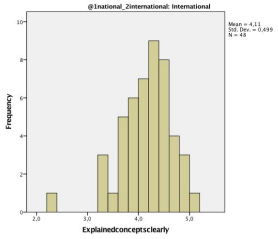
(i) Engaged students National



(j) Engaged students International

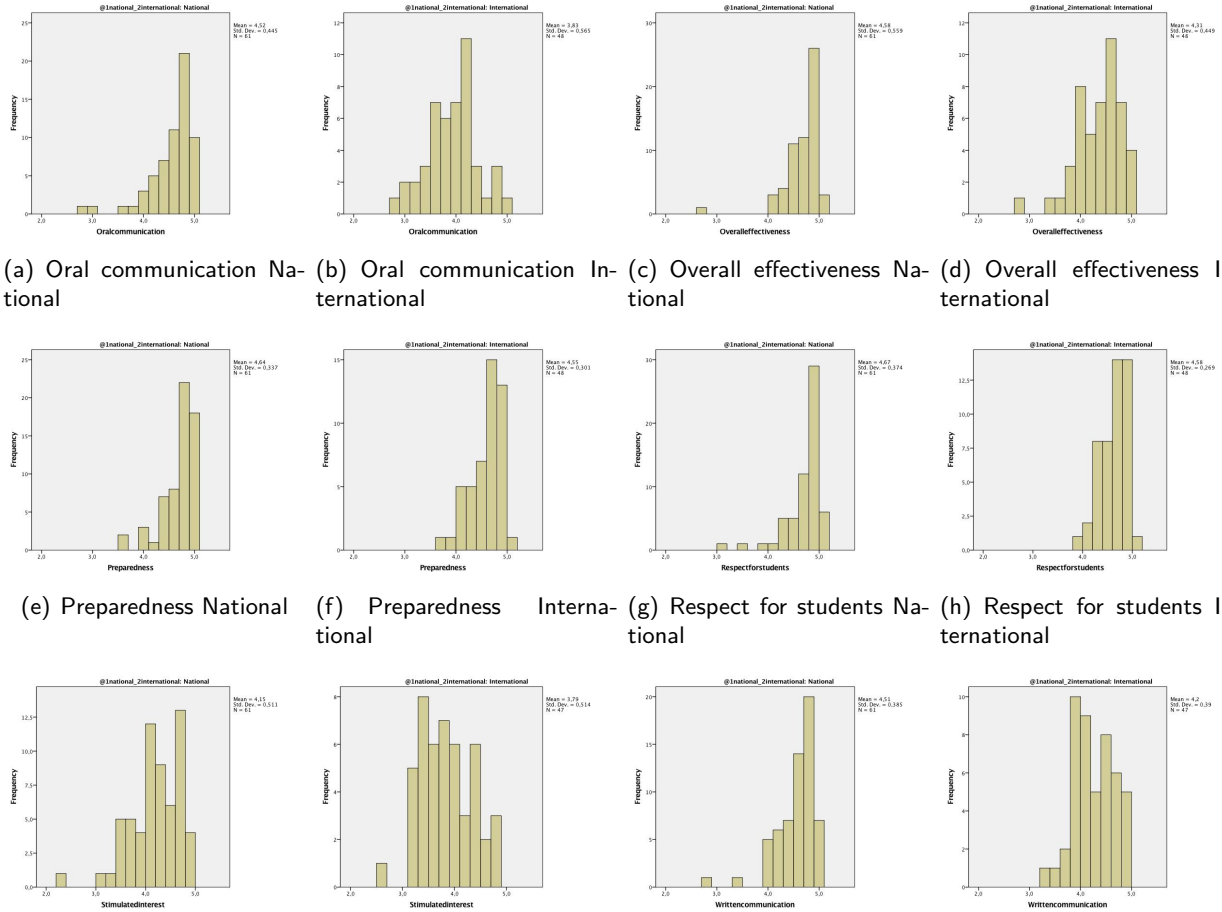


(k) Explained concepts clearly National



(l) Explained concepts clearly International

Figure 13: Histograms of the distributions of the national and international GTAs for the spring semester of teaching.



(a) Oral communication National

(b) Oral communication International

(c) Overall effectiveness National

(d) Overall effectiveness International

(e) Preparedness National

(f) Preparedness International

(g) Respect for students National

(h) Respect for students International

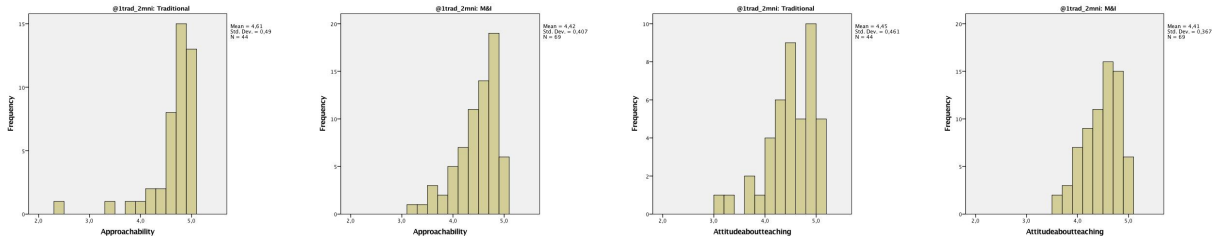
(i) Stimulated interest National

(j) Stimulated interest International

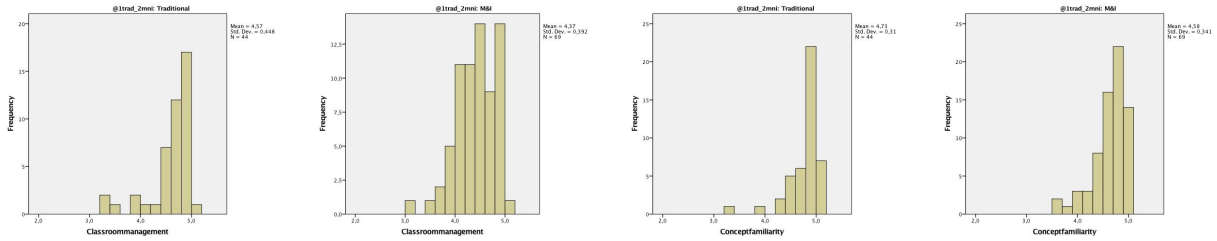
(k) Written communication National

(l) Written communication International

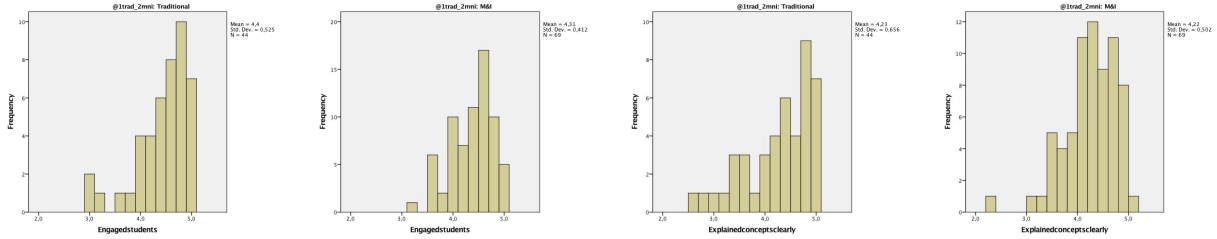
Figure 14: Histograms of the distributions of the national and international GTAs for the spring semester of teaching.



(a) Approachability Traditional (b) Approachability M&I (c) Attitude about teaching Traditional (d) Attitude about teaching M&I

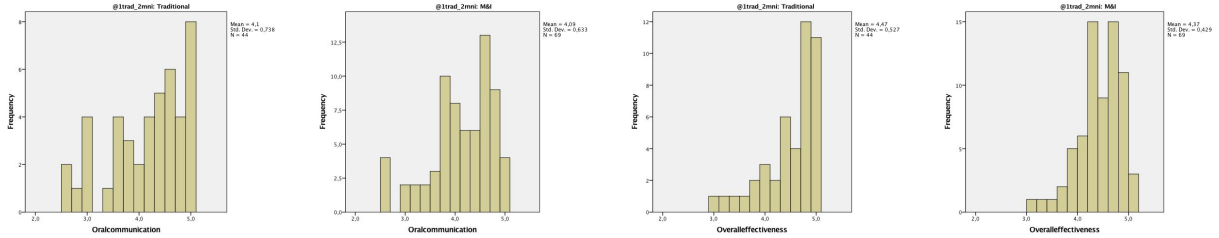


(e) Classroom management Traditional (f) Classroom management M&I (g) Concept familiarity Traditional (h) Concept familiarity M&I

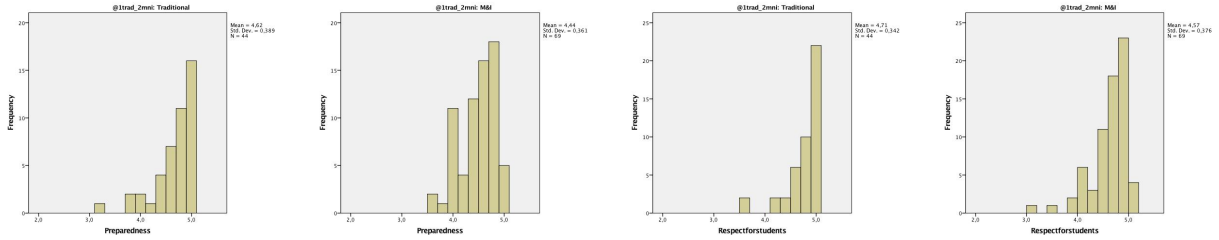


(i) Engaged students Traditional (j) Engaged students M&I (k) Explained concepts clearly Traditional (l) Explained concepts clearly M&I

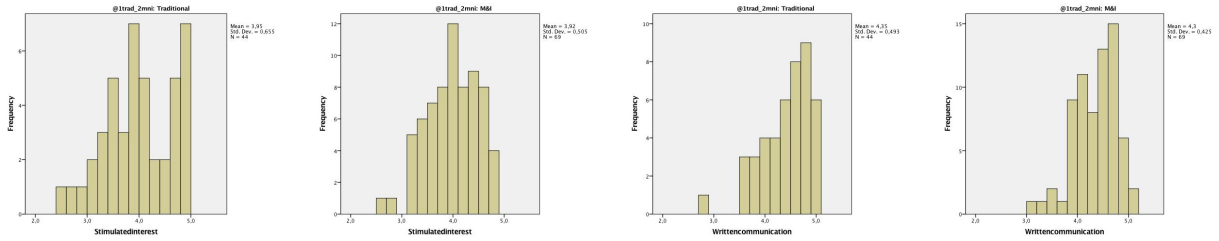
Figure 15: Histograms of the distributions of the traditional and M&I GTAs for the fall semester of teaching.



(a) Oral communication Traditional (b) Oral communication M&I (c) Overall effectiveness Traditional (d) Overall effectiveness M&I

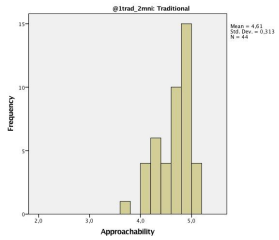


(e) Preparedness Traditional (f) Preparedness M&I (g) Respect for students Traditional (h) Respect for students M&I

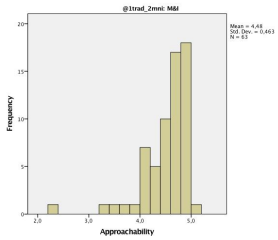


(i) Stimulated interest Traditional (j) Stimulated interest M&I (k) Written communication Traditional (l) Written communication M&I

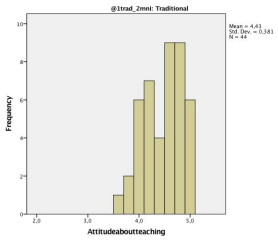
Figure 16: Histograms of the distributions of the traditional and M&I GTAs for the fall semester of teaching.



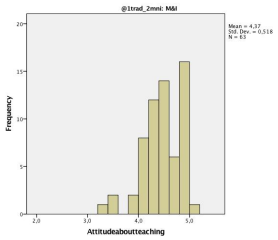
(a) Approachability Traditional



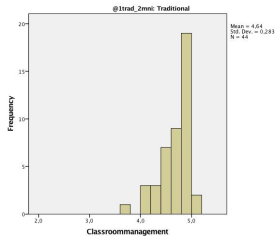
(b) Approachability M&I



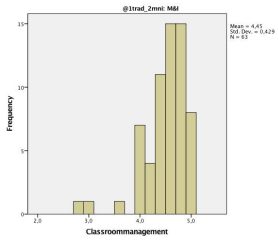
(c) Attitude about teaching Traditional



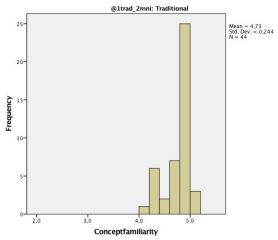
(d) Attitude about teaching M&I



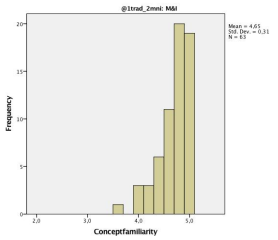
(e) Classroom management Traditional



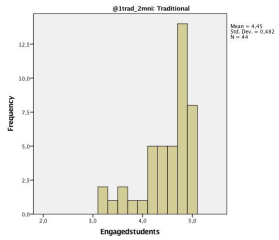
(f) Classroom management M&I



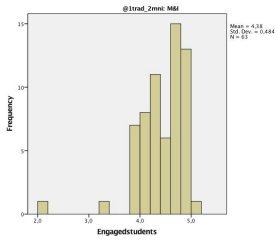
(g) Concept familiarity Traditional



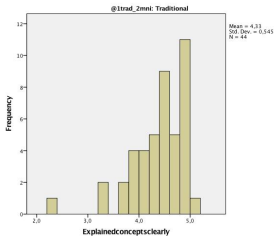
(h) Concept familiarity M&I



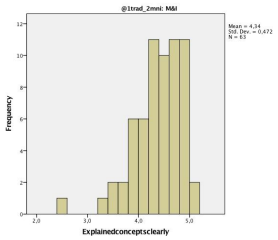
(i) Engaged students Traditional



(j) Engaged students M&I

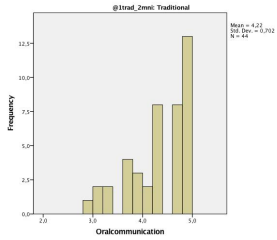


(k) Explained concepts clearly Traditional

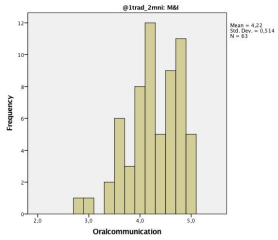


(l) Explained concepts clearly M&I

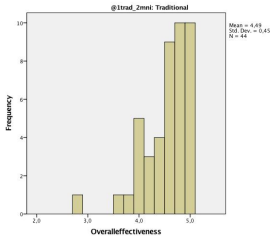
Figure 17: Histograms of the distributions of the traditional and M&I GTAs for the spring semester of teaching.



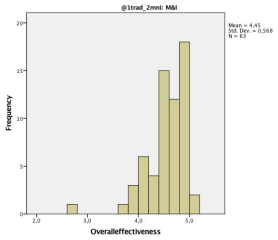
(a) Oral communication Traditional



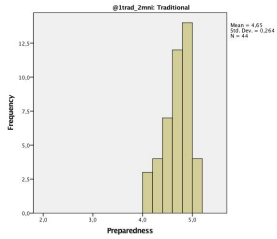
(b) Oral communication M&I



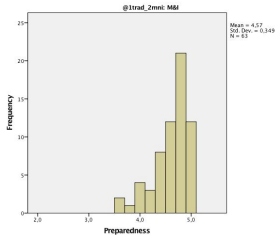
(c) Overall effectiveness Traditional



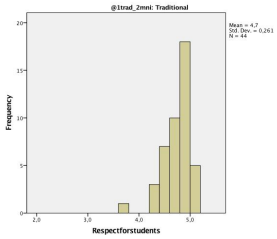
(d) Overall effectiveness M&I



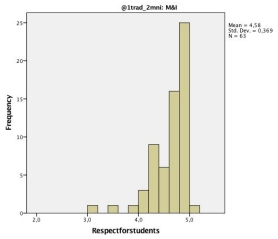
(e) Preparedness Traditional



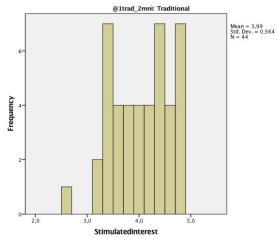
(f) Preparedness M&I



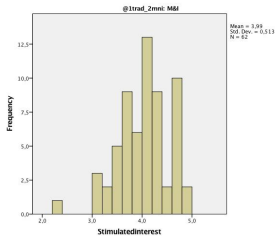
(g) Respect for students Traditional



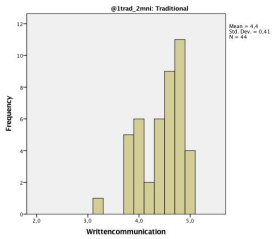
(h) Respect for students M&I



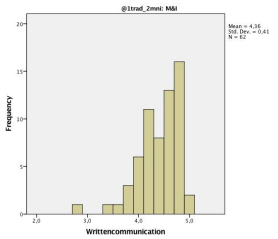
(i) Stimulated interest Traditional



(j) Stimulated interest M&I



(k) Written communication Traditional



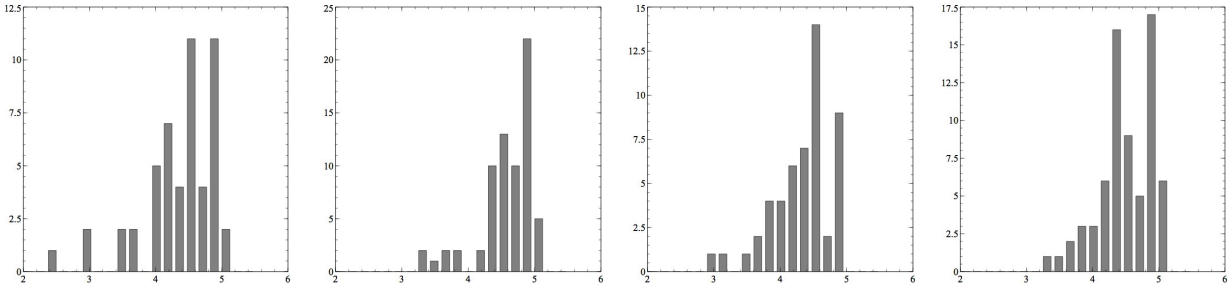
(l) Written communication M&I

Figure 18: Histograms of the distributions of the traditional and M&I GTAs for the spring semester of teaching.

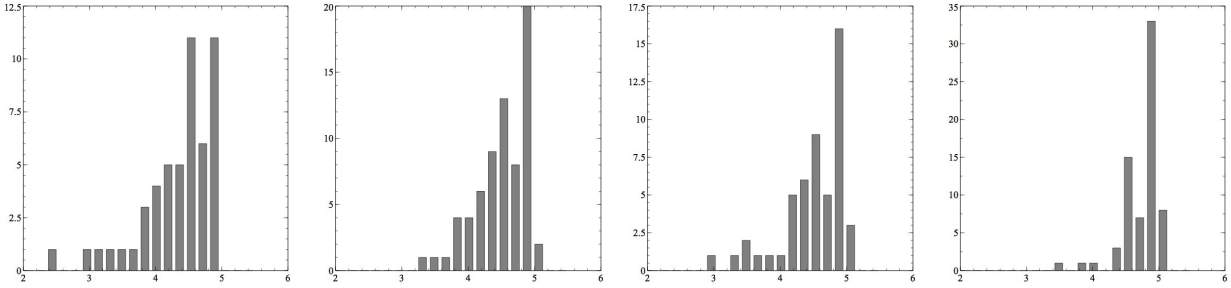
A.2 Mann-Whitney test by pre and post groups assumption

In this section we are comparing the shapes of the pre and post groups for every question of the TAOS survey, in order to see that the distributions have similar shapes. To do that we represent the histograms of the pre and post groups for the fall and spring semester of teaching. The histograms are shown in the figures 19, 20, 21 and 22.

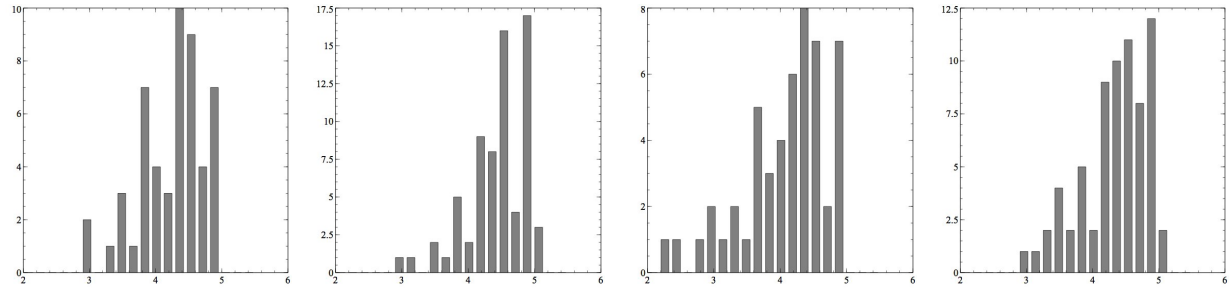
As we can see in all the figures from 11 to 22, although the shapes are not exactly equal, they have enough similarities to apply the Mann-Whitney test in order to compare the difference in the means of the distributions.



(a) Approachability Pre (b) Approachability Post (c) Attitude about teaching Pre (d) Attitude about teaching Post

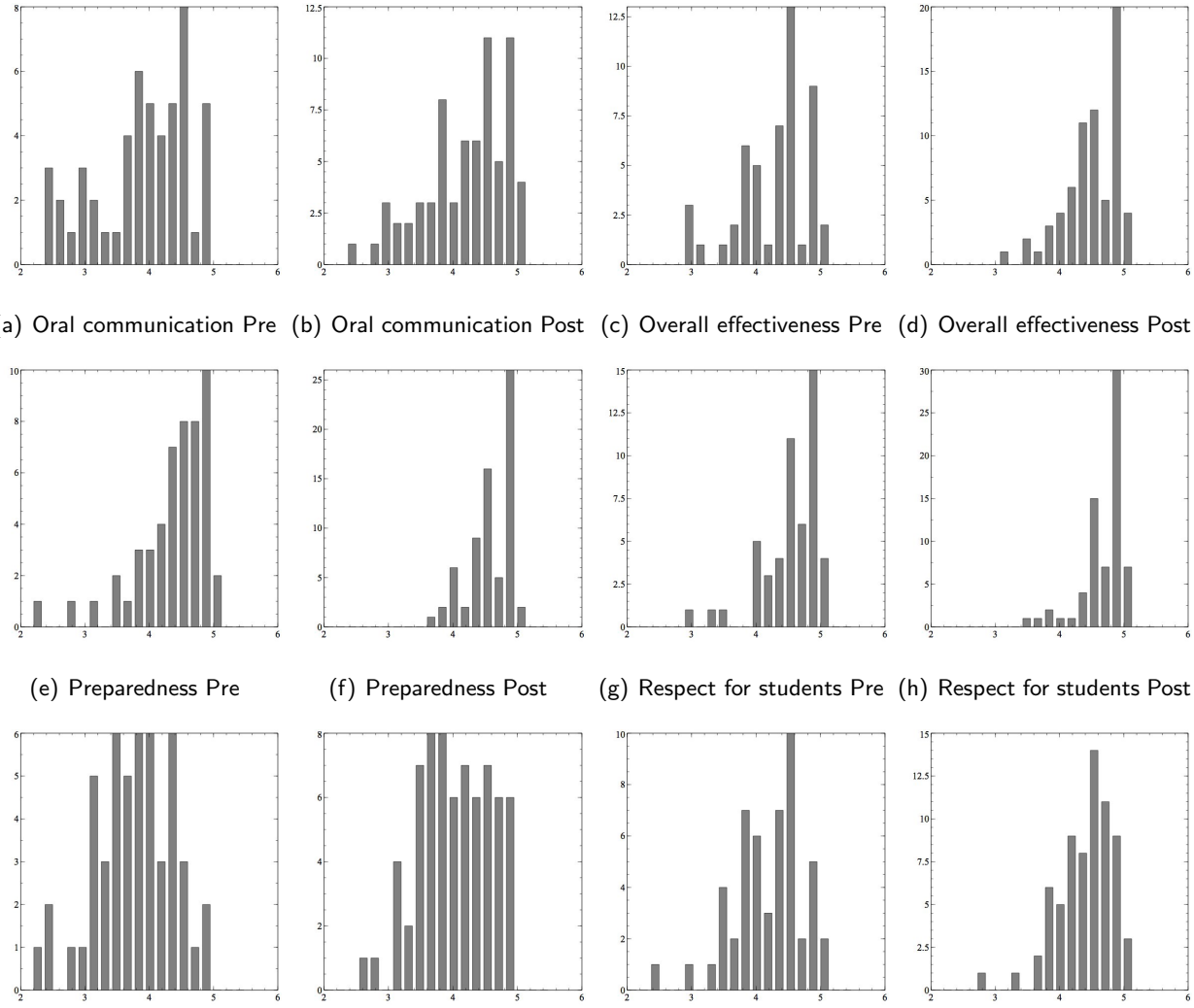


(e) Classroom management Pre (f) Classroom management Post (g) Concept familiarity Pre (h) Concept familiarity Post



(i) Engaged students Pre (j) Engaged students Post (k) Explained concepts clearly Pre (l) Explained concepts clearly Post

Figure 19: Histograms of the distributions of the pre and post groups of GTAs for the fall semester of teaching.

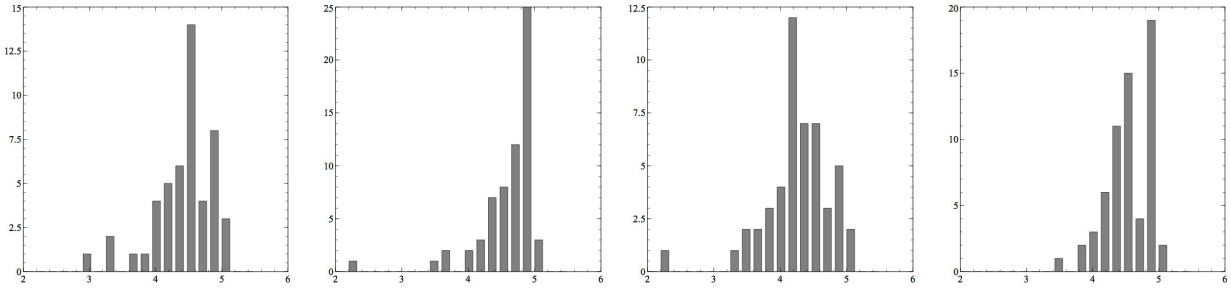


(a) Oral communication Pre (b) Oral communication Post (c) Overall effectiveness Pre (d) Overall effectiveness Post

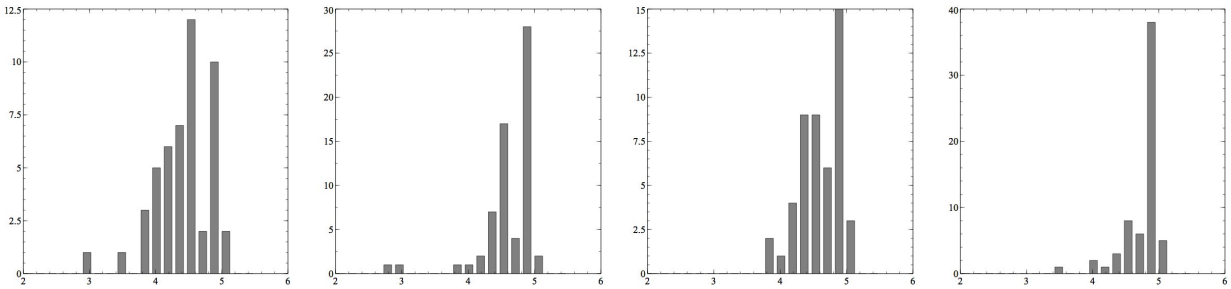
(e) Preparedness Pre (f) Preparedness Post (g) Respect for students Pre (h) Respect for students Post

(i) Stimulated interest Pre (j) Stimulated interest Post (k) Written communication Pre (l) Written communication Post

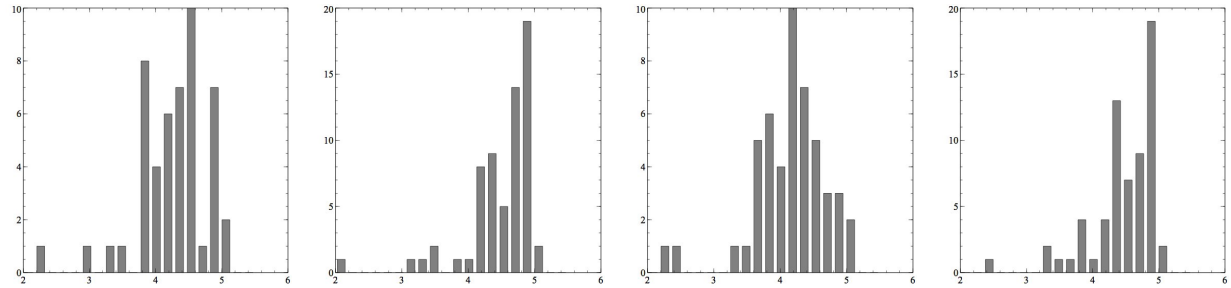
Figure 20: Histograms of the distributions of the pre and post groups of GTAs for the fall semester of teaching.



(a) Approachability Pre (b) Approachability Post (c) Attitude about teaching Pre (d) Attitude about teaching Post

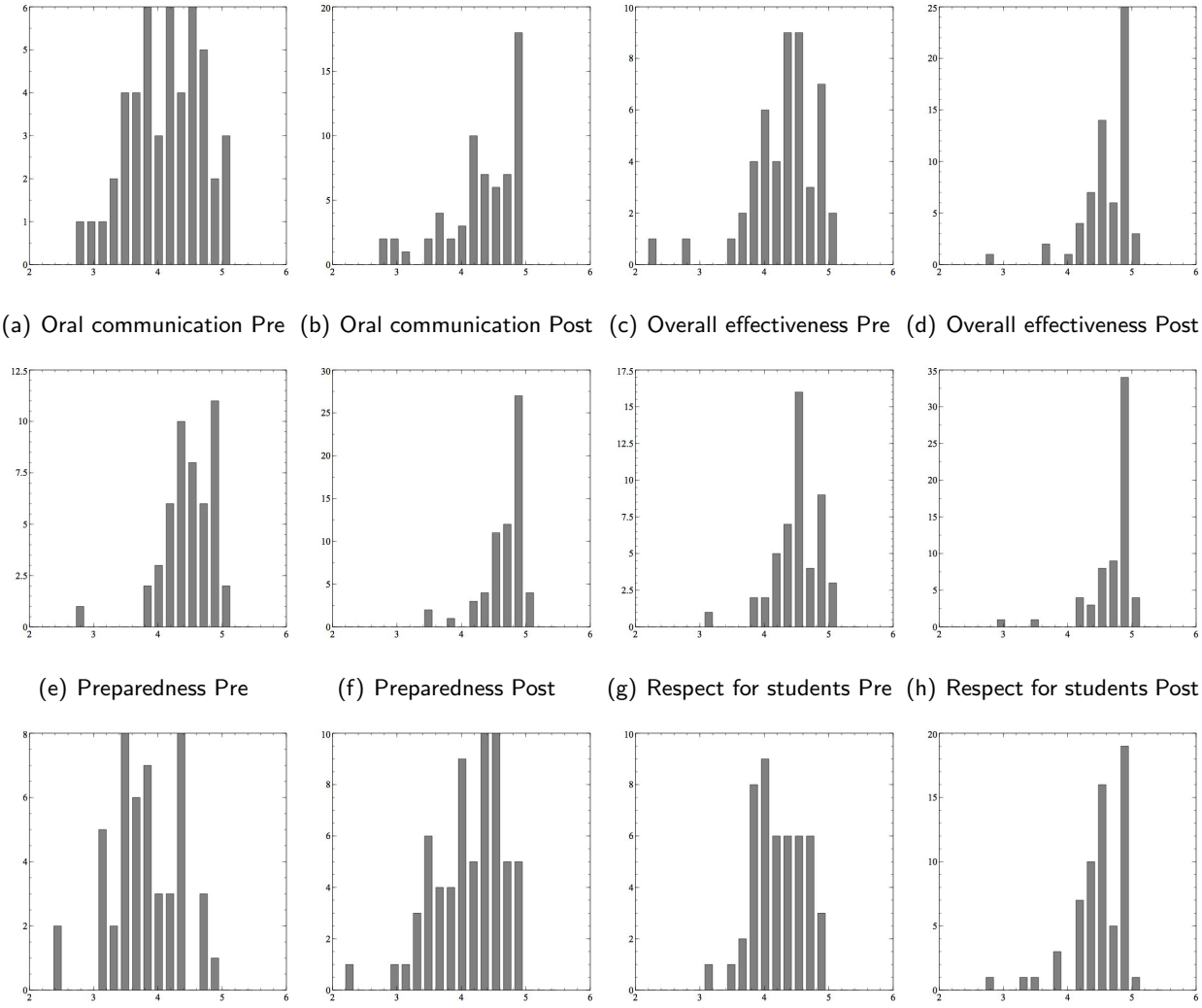


(e) Classroom management Pre (f) Classroom management Post (g) Concept familiarity Pre (h) Concept familiarity Post



(i) Engaged students Pre (j) Engaged students Post (k) Explained concepts clearly Pre (l) Explained concepts clearly Post

Figure 21: Histograms of the distributions of the pre and post groups of GTAs for the spring semester of teaching.



(a) Oral communication Pre (b) Oral communication Post (c) Overall effectiveness Pre (d) Overall effectiveness Post

(e) Preparedness Pre (f) Preparedness Post (g) Respect for students Pre (h) Respect for students Post

(i) Stimulated interest Pre (j) Stimulated interest Post (k) Written communication Pre (l) Written communication Post

Figure 22: Histograms of the distributions of the pre and post groups of GTAs for the spring semester of teaching.